

Surviving misconduct is one thing, accountability is another

Recent high-profile research misconduct cases underline an urgent need for a more formal approach to the way in which laboratories organize and conduct their research.

Well over a million papers were published last year in journals whose contents are recorded by the Institute of Scientific Information. Against that level of activity, tens of incidences of fraud and other severe research misconduct around the world is a statistically small number. But the implications of such cases for individual institutions, and even for a national scientific community, can be extraordinarily costly and damaging — in time, in possible litigation, in the smearing of colleagues' reputations, and in public confidence in science. That is why, for example, recent cases of fraud at the laboratories of the prestigious French institution INSERM and of the equally prestigious Max Planck Society have proved so significant and painful.

Acting promptly, decisively and transparently is what has not happened at INSERM but what, commendably, did happen in the Max Planck Society's investigations into a recent prominent case of fraud — to the credit of Hubert Markl, head of the Max Planck Society, but also to the credit of the Deutsche Forschungsgemeinschaft (DFG), which has provided one of the best overviews of misconduct and how to deal with it (see www.dfg.de). It is encouraging that other European agencies, including INSERM, are now collaborating with their German counterparts and with others who have already set standards — in Scandinavia and Britain, for example.

Following principles adopted in the United States, the DFG's document includes the threat that, in future, an institution lacking adequate procedures for handling misconduct might not receive funds. But pre-empting misconduct is just as important, and the DFG is therefore to be commended for hinting that the same penalty could apply to those institutions that lack good guidelines for the proper conduct of research itself. One prosaic but essential possibility is an insistence that institutions keep good laboratory notebooks and also archive them for a guaranteed ten years — such notebooks are often

crucial in understanding what has gone wrong.

In any country, there will inevitably be both explicit and implicit attempts to undermine such goals. Explicit, because many laboratories are run by people who are under great competitive pressure and claim that they have neither the time nor the inclination to introduce new approaches to the management of their colleagues. Implicit because in biomedical research, where extreme misconduct tends to occur most frequently, funding is increasing — to an extraordinary extent in the case of the US National Institutes of Health (see page 734). Such a beneficent environment is in itself not conducive to tightening up the quality of one's act, unless the funding agencies insist on it as a condition for receiving support.

Much of this comes down to an issue of leadership and of institutional accountability for the use of public funds. Regrettably or not, many academic researchers find themselves required to achieve institutional targets — long gone, for them, are the relaxed days of disinterested enquiry for its own sake. But too often, a sense of direction and accountability has not trickled down the lab hierarchy. Younger scientists are still left in isolation and without much supervision or feedback, and the principles of good person-management — appraisal and career development — are frequently in short supply.

In short, like it or not, the pressure of misconduct cases coming from one direction, and a greater sense of institutional purpose coming from the other, are forcing researchers to act in a way that requires accountability to be more formalized, in both research and employment practice, within laboratories. The question is whether institutions are willing to grapple with the commitment this requires in both time and staff, and whether funding agencies will be prepared to contemplate the additional costs. Presumably, in the absence of enforcement from above, they will not. □

Fanning the winds of change

Persuading countries to increase their investment in research often means starting at the bottom.

The tourists who flock to the sun-drenched beaches of the Caribbean can easily forget the harsh realities that face those who live on the islands. Television pictures of last month's hurricane brought a little of this reality, albeit briefly, into our sitting-rooms. But they still failed to show the poor housing and lack of adequate health care that are part of everyday life. Meanwhile, too many of the region's brightest scientists, already few enough in number, continue to leave to work in the United States or elsewhere.

A number of attempts have been made to address this situation. One, modest enough in its own way, was the creation ten years ago of a Caribbean Academy of Sciences, in a bid both to raise the status of science in the region and encourage a closer working relationship between researchers on different islands. An impressive programme of activities has been established (see page 735). But the academy still

has a long way to go to achieve its objectives. High among the obstacles is the lack of an indigenous scientific culture in the Caribbean. This is partly the result of a geographical location that has allowed the region to survive on its natural resources, rather than requiring it to survive on its wits. The legacy is a low level of investment in research and development that few outside the scientific community see the need to change.

But change is essential if the region is to thrive in a global, knowledge-based economy. And it needs to begin in the schools and on the streets. Imaginative solutions are required. One suggestion heard at the academy's tenth anniversary meeting, for example, was to tap into the curiosity of the young by exploring the science tacitly involved in fabricating instruments for the ubiquitous steel bands. A more productive meeting of the two cultures is difficult to imagine. □



French researchers kick up a storm over CNRS 'reforms'

[PARIS] Draft plans for a major overhaul of France's Centre National de la Recherche Scientifique (CNRS), Europe's largest fundamental research agency, have run into a storm of protest from labour unions representing scientific staff.

The agency, which has 27,000 staff, currently operates primarily through its own nationwide network of laboratories. The government's proposals put it on the road to becoming a funding agency for university-based research. The changes would include giving universities joint responsibility for the running of CNRS laboratories.

CNRS funds would be opened up to bidding from competing university research groups. The Ministry of Research argues that such changes, which were outlined to the board of directors of CNRS two weeks ago, will lead to greater efficiency in the way research is carried out.

But the unions have condemned the pro-

posed changes as a recipe for "dismantling" the agency. This view appears to have the tacit backing of many top-level CNRS managers. Criticizing the lack of consultation with other interested parties, the unions have promised that they will fight the reforms "to the bitter end".

The unions' reaction is based on a leaked draft of proposals that have been drawn up by Edouard Brezin, a physicist at the Ecole Normale Supérieure in Paris, who is president of the board of CNRS.

An apparent desire to shift control of research from the CNRS to the universities was reflected in the text by proposed changes to the agency's statutes that would place it not only under the jurisdiction of the "ministry responsible for research," as in the past, but also under the ministry responsible for "higher education".

And, rather than keeping its powers to "create" laboratories and manage research



nationally, the text describes the agency's new role as "recognizing and subsidizing" research entities within universities, and other bodies including private companies.

The agency would no longer exert full control over its laboratories, but would exercise joint responsibility for them with the laboratory's parent university. Its responsibility for constructing and managing big science facilities would also disappear. This function, according to ministry sources, would be assumed by the ministry itself.

The chairpersons of six sections of the CNRS's national committee — the body that evaluates all CNRS laboratories and shapes the agency's research priorities — have said they will resign unless the draft decree is revoked, arguing that it threatens to damage "the whole of French research".

The unions' argument that the changes would turn CNRS into little more than a research council is disputed by both Brezin and Vincent Courtillot, principal adviser to Claude Allègre, the minister for national education, research and technology.

They argue that the plan merely represents a further step in a process of closer integration of the agency into the university system that has been under way ever since the first CNRS laboratories associated with universities were created in 1966.

Brezin and Courtillot assert that the text is preliminary and open to debate. Indeed, late last week Brezin issued a revised text, taking into account some of the concerns expressed by the unions and researchers. "The draft text was not written in stone," says Brezin.

Courtillot argues that, although "the unions would prefer to talk at length about general problems first, we think it is better to show a text and get reactions," adding that

Universities to get more control of biomedicine

[PARIS] A reform of the French national biomedical agency INSERM that was abandoned during the summer is set to be resuscitated by the government, with the release of a new decree expected by the end of the year.

The old plan met opposition from unions representing researchers (see *Nature* **395**, 630; 1998).

The new reforms, which are also likely to meet opposition, mirror those proposed for the Centre National de la Recherche Scientifique (see above), with INSERM having to share control of its laboratories with the universities. A formal contract between INSERM, the Ministry of National Education, Research and Technology, and the Conference of University Vice-chancellors will set out the common strategic goals of joint groups of university and INSERM researchers.

The new arrangements will mean an influx of university researchers into INSERM laboratories, which



Griscelli: wants to open doors for university researchers.

are at present independent of universities. Around 80 university laboratories will also be given INSERM status over the next five years, according to Claude Griscelli, the agency's director general.

Griscelli says that the new reform will integrate the agency's research activities into a broad national strategy for the life sciences, being drafted by a ministerial committee, that will include all the research organizations and universities.

Ultimately, Griscelli sees INSERM being built around the so-called Instituts

Fédératifs de Recherche (IFR). These create a critical mass in particular scientific areas by bringing together local scientists from the various research agencies, hospitals and universities, forming loose federations of laboratories.

Griscelli plans for around 15 of the 70 IFRs so far created to become full-blown research institutes. Each would have around 300 staff, and would share large facilities, such as transgenic knock-out mice laboratories. Another layer of IFR's would form national networks in particular research areas such as ageing and ophthalmology.

Griscelli also wants to shake up research funding by shifting financing from laboratories to grants awarded to individual groups, based on competitive proposals.

INSERM's research itself will in future be oriented towards medical applications of genome research, according to Griscelli. **D.B.**

"we really are open and will listen to changes".

One initial proposal that has already been modified is that CNRS would be deprived of the authority to create laboratories, such as those on its campuses at Gif-sur-Yvette on the outskirts of Paris. The revised decree allows for the creation of CNRS laboratories, but these would be the exception. "We are not against CNRS opening new labs, but as a general rule it should take place in discussions with universities," says Courtillot.

The major point of contention over the draft decree concerns the balance of forces between the agency and the universities, a topic with a long history. The agency was created in 1939 to compensate for the weak research base in the universities. Indeed, today the only French universities carrying out international-level research are those with large teams of CNRS researchers.

Many researchers say that university science is still weak, and that it would be premature to increase their power over CNRS-financed research. Observers argue that French universities tend to put student and local needs first, and cannot properly define or evaluate national research strategies.

The vice-chancellors and the scientific bodies within universities are elected, and this democracy is said to generate a lack of competitiveness. In contrast, the CNRS has rigorous evaluation mechanisms and a reputation for putting scientific excellence first.

A reform of the university system is a prerequisite for any profound reform of CNRS, assert several observers. They argue that, given that university researchers are civil servants, a university laboratory winning a grant would not be freely able to hire researchers, making the idea of a research council meaningless within a French context.

Courtillot agrees that it would be premature to transform CNRS into a research council along the lines of the US National Science Foundation or the UK research councils. But he says the government wants the CNRS to move in this direction, and that the universities should ultimately be the major players in French research.

Courtillot concedes that many universities are not yet capable of taking over responsibilities from the agency, but says that others are ready, and that more will be in time. "In 1988, less than ten of the then 80 French universities were in a position to have their own research policies, but now 40 of the 100 universities are able to do so," he says.

Another ministry proposal is to replace the current system of funding laboratories with one in which individual teams would be financed on the basis of competitive proposals. The idea, says Courtillot, is to blur the boundaries between the institutions, with teams from different research agencies and universities uniting to form new entities and to seek joint project funding. **Declan Butler**

Japan's universities resist plan for greater autonomy

[TOKYO] A bid by the Japanese government to transform national universities into semi-autonomous institutions is meeting strong resistance both from leading academics and from the Ministry of Education, Science, Sports and Culture (Monbusho).

Last December, a proposal to change the status of the nation's two leading universities — Tokyo University and Kyoto University — was omitted from the final version of the government's administrative reform plan (see *Nature* 389, 897; 1997). But it has resurfaced since the appointment of the new prime minister, Keizo Obuchi, who has pledged to carry out drastic reorganization of Japan's administration.

The administrative reform plan, which originally targeted national research institutes attached to government ministries and agencies (see *Nature* 395, 211; 1998), will now also target some of the national universities — including Tokyo and Kyoto.

But the plans are running into strong opposition. Last week Leo Esaki, the Nobel prizewinner and former president of Tsukuba University, said that increased autonomy and drastic changes in their management systems could have negative effects on the universities.

"Given the Japanese universities' lack of competitiveness and management skills, turning them into semi-autonomous bodies with administrative independence will only make things chaotic," said Esaki. He argued



Esaki: warning of negative effects.

that universities needed instead to begin reforms at a fundamental level, such as introducing a proper system of peer review.

Meanwhile, the Council on University Education, chaired by Akito Arima, the education minister and former president of Tokyo University, last week announced plans to set up an independent body to assess the performance of national universities.

The move is seen as a bid to resist the government's plans, including external evaluation of the universities' performance every three to five years, by indicating the universities' willingness to increase their effectiveness without the government's restructuring. They argue it will damage both education and research.

Although the government hopes to finalize its reorganization plans by next January, many predict that Monbusho and the universities will not give in, and may try to generate sufficient political opposition for the government to back down.

But some university researchers support the plans, arguing that the national universities' aversion to greater autonomy hinders the spread of venture businesses and industry/university collaboration. **Asako Saegusa**

US Congress rebuffs data copyright law

[WASHINGTON] Critics of proposed legislation that would assign sweeping copyright protection to commercial online databases scored a victory last week when the US Congress passed a bill with the controversial measure stripped out.

Organizations including the American Association for the Advancement of Science and the Association of Research Libraries (ARL) and the Institute of Electrical and Electronics Engineers had been lobbying for months against the legislation, which was sponsored by Representative Howard Coble (Republican, North Carolina) and Senator Rod Grams (Republican, Minnesota).

The proposed change in copyright law, its opponents say, would stifle the flow of scientific information by granting broad and vaguely defined proprietary rights to data that are now exchanged freely (see *Nature* 394, 410; 1998).

When the measure moved easily through the House of Representatives in May and

was then folded into a larger copyright bill on a fast track for passage, "the visibility of the issue was heightened," according to Prudence Adler of the ARL. An intensive letter-writing and lobbying campaign led to House and Senate negotiators dropping the database provision from the larger copyright bill earlier this month.

Around 50 companies and associations, representing a range of political views and interests, weighed in against the database clause. Several federal agencies also pointed out potential constitutional problems.

By the time of the vote, 15 senators urged that the database clause be dropped because it had not been debated thoroughly and serious disagreements remained.

But Judiciary Committee chairman Orrin Hatch (Republican, Utah) and other Republican leaders promised the bill's sponsors that the database issue will get an early hearing when a new Congress returns to Washington in January. **Tony Reichhardt**



Rallying call: Washington's first annual cancer march last month demanded more cancer research.

US cancer body calls for ideas from all disciplines

[WASHINGTON] Scientists who may never have thought about applying their ideas to health research are being asked by the US National Cancer Institute (NCI) to come up with innovative technologies to support cancer research and treatment.

Under a programme to start early in the new year, NCI plans to spend \$48 million over five years on the 'unconventional innovations'. It hopes, for example, to find novel approaches for identifying molecular alterations in the body from a distance, potentially opening up an era of early and non-invasive cancer detection and treatment.

NCI has advertised the programme in *Chemical and Engineering News* and *Physics Today*, seeking ideas for solving cancer research problems from chemists, engineers, physicists and materials scientists who have had no association with health research.

The programme is part of a wider trend at the National Institutes of Health (NIH) that is likely to accelerate as it spends the \$2 billion budget windfall it received from Congress and President Bill Clinton last week (see page 734).

NIH institutes — of which NCI is the largest — are exploring several avenues for engaging non-biologists more fully in research to improve public health.

"We're being very open-minded right now" says Carol Dahl, director of NCI's office of technology and industrial relations, which is administering it. "We want to get people to think about the problem. We don't yet have a complete conception of all the technologies that could contribute."

The initial call for ideas, which is open until the end of November, will help Dahl's office to frame a request for contracts which will be issued early in 1999.

The programme will look for imaging technology that could enable physicians to detect cancer in patients at a very early stage.

Genomic research at NCI is expected to identify molecular signatures for cancer. Then, imaging technology that could find such signatures in the body might allow the disease to be diagnosed long before it develops into visible tumours.

Achieving this "will require the development and integration of a series of capabilities including highly specific molecular recognition, signalling capability, controllable intervention capabilities, methods for monitoring intervention release and impact, and biotolerance," says NCI's announcement of the programme, posted on the Web (<http://amb.nci.nih.gov/uip.htm>).

Dahl anticipates supporting multi-disciplinary teams, bringing university scientists together with industrial partners that would be ready to commercialize the results. She plans to break with standard practice at NIH by issuing contracts rather than grants, so the researchers' work will be coordinated by staff at NCI and directed towards the goal of developing working technologies.

"By having contracts, we'll have more contact with researchers, as we try to bring together a community of scientists to work in this area," she explains.

In using contracts and imploring researchers to think "outside the box" for their own sub-discipline, the initiative appears to borrow the approach to supporting research taken by the Defense Advanced Research Projects Agency (DARPA).

DARPA used to operate chiefly in physics, engineering and computer science, but has recently extended into biology. Its approach has been praised for its results, but criticized by some scientists as over-prescriptive. Dahl says NCI looked at both DARPA and the Advanced Technology Programme at the National Institute of Science and Technology, and "attempted to extract the best aspects" of both.

Colin Macilwain

Proposed restrictions relaxed on research on mentally disabled

[WASHINGTON] The National Bioethics Advisory Commission (NBAC) is this week expected to endorse a final report on the protection of mentally disabled research subjects. The report is likely to be slightly less restrictive than a draft version circulated in July that was roundly criticized by US medical schools.

The report is part of a process that is likely to lead to new rules and restrictions on the use of such subjects in clinical research. But representatives of the medical schools have argued that it could inhibit vital research into Alzheimer's disease and other ailments (see *Nature* 394, 713; 1998).

According to Eric Meslin, executive director of NBAC, the final report will probably "broaden the discretion" of the Institutional Review Boards to permit experiments that involve some risk to patients but will not benefit them directly.

It will allow review boards to ask for a waiver from the Department of Health and Human Services to allow such research to proceed. Previous proposals to tighten control on research involving mental patients have been successfully resisted by researchers.

Meslin also says the commission will acknowledge that other groups of individuals whose ability to give informed consent is impaired, and not just the mentally ill, could be subject to the proposed extra protections. This would defuse the argument that the rules would single out psychiatric research for special scrutiny.

The final report, says Meslin, is likely to ease the requirement, proposed in the draft, that researchers carry out a "capacity assessment" of research subjects before proceeding with an experiment. The requirement will now apply only if the experiment posed more than minimal risk for the subjects, he adds.

Meslin anticipates continued resistance from researchers to the commission's recommendations. "Some of them will be acceptable and some of them will be difficult to accept," he says.

The NBAC report will be formally submitted to the National Science and Technology Council, a body made up of the government's top science administrators, which will decide what to do to implement it. "We'll also encourage voluntary compliance by researchers and Institutional Review Boards," says Meslin.

NBAC was established by President Bill Clinton in 1995 in response to several scandals concerning the treatment of human research subjects. The general performance of the review boards will be the subject of a second report from the Commission, due to be published in May 1999.

Colin Macilwain

Europe's space funds feel the squeeze ...

[MUNICH] Space scientists in Europe are trying to convince member states of the European Space Agency (ESA), whose space ministers are due to meet next month, to lift a three-year funding 'cap' on the agency's science programme.

The cap has not only significantly eroded the programme, but also helped to plunge it into its most serious financial crisis. According to Roger Bonnet, ESA's space science director, if the cap is not lifted, some programmes in ESA's long-term Horizons 2000 science plan may have to be abandoned.

At present, ESA's member states are divided roughly equally in their willingness to lift the cap. But since a unanimous decision is required, the cap is likely to remain.

Horizons 2000 includes four approved large 'cornerstone' missions, each costing about ECU650 million (US\$786 million), balanced by medium-sized missions at about half this cost in various disciplines.

The programme was approved by ESA in 1994 when a continuous growth in budget was foreseen. Two years later, however, the space ministers, facing economic difficulties at home, decided to hold the budget constant for three years. There was to be no increase for inflation, and this would automatically be extended for a further two years unless ministers agreed unanimously to raise it again.

ESA has since introduced extensive cost-cutting and efficiency measures. It has also been dogged by bad luck. In particular, two important Solar System missions — ESA's Cluster and Russia's Mars 96, with many ESA instruments on board — were lost on launch in 1996 (see *Nature* **381**, 541; 1996).

To maintain a Horizons programme serving all scientific disciplines, Bonnet decided, with the approval of the ESA council, to relaunch Cluster with existing spares, and to



To Mars? The unforeseen Mars Express weighs heavily on ESA's already stretched purse-strings.

develop the Mars Express concept.

The second of these would carry duplicates of the Mars 96 instruments, and complement the flotilla of spacecraft that the US space agency NASA is sending to Mars in the next few years. The cost of Mars Express has been capped at ECU150 million.

But such unplanned-for expenses have placed further strain on ESA's budget. And more trouble could be ahead. The launch of the X-ray astronomy cornerstone mission XMM, scheduled for 1999, has been delayed for six months by technical problems.

The delay will add to the costs because of the need to keep teams together, while uncertainty hangs over Russia's ability to make good on its promise to provide a free Proton launch for the gamma-ray astronomy mission Integral, scheduled for launch in 2001.

An alternative Ariane-5 launch remains possible, but would cost ECU150 million. Also, ESA's new policy of insuring science missions will probably increase some launch costs. Insurance of ECU18 million is likely to be taken out for the 2003 Cluster relaunch.

Budget constraints have strained relation-

ships between Solar System scientists and astrophysicists, who until now have worked relatively happily together within ESA's consensus-driven system of programming.

To make further savings to accommodate Solar System missions, Bonnet has suggested combining two astrophysics missions that are both in the design phase: the infrared cornerstone mission FIRST, and the medium-sized microwave background mission Planck, both approved in 1997.

Bonnet's original plan to place both missions in a single container on the same launch vehicle was rejected in the summer by ESA's science programme committee as restricting the science of each mission (see *Nature* **387**, 639; 1997).

ESA is now assessing the less radical option of keeping the two missions physically separate on the same launch, and delaying the launch to 2007 — four years after the scheduled Planck launch and two years after FIRST's scheduled launch — in order to spread costs.

The merger debate caused uproar in the astrophysics community and triggered a wave of discontent. "The scientific community was alarmed by the single-minded effort to save money through complete merging of two missions with different purposes," says Reinhard Genzel, a director of the Max Planck Institute for Extraterrestrial Physics in Garching, and principle investigator for one of the Planck instruments. "It was obvious to everyone what a high risk this would be."

Pleased that a complete merger has been rejected, Genzel still feels that any delay forced on Planck "is unfair," because both Planck and FIRST were selected through ESA's normal mechanism of peer-review competition, while Mars Express "came in the back door".

"ESA procedures are breaking down," he says, pointing out that Mars Express was originally approved on the understanding that it would not affect the rest of the Horizons programme.

Bonnet says he will do his best to keep Mars Express on track, as he is determined to maintain a programme balanced between space science disciplines. But he emphasizes that no fully approved Horizons 2000 mission will be abandoned.

Scientists at ESTEC, ESA's scientific and technical arm, recently passed a vote of no confidence in the science directorate's management, having become frustrated by the lack of decision-making.

Only a budget increase would diffuse the tensions and allow all disciplines to be satisfied. But reestablishing a link with inflation in next month's 1999 budget discussions seems unlikely, as too many member states remain set against it.

Alison Abbott

... while UK scientists seek private backing

[MUNICH] British space scientists are making an urgent appeal to private funders to support a small equipment package that would land on the surface of Mars as part of the Mars Express mission.

The Mars Express spacecraft was originally developed to carry instruments duplicating those lost when the Russian Mars 96 mission, intended to analyse the martian atmosphere, failed on launch in 1996. At an early stage, however, its scope was

broadened to include a lander to transmit data on the atmosphere and the planet's surface back to Earth.

The UK-led lander, called Beagle, has been approved for inclusion, but only if its principal investigator, Colin Pillinger, director of the Planetary Sciences Research Institute at the Open University in Milton Keynes, can raise private funding.

Estimated to cost around £25 million (US\$43 million), Beagle would look in particular for evidence of life or former life on the planet.

But its prospects appear bleak. The Particle Physics and Astronomy Research Council has said it cannot afford to contribute, and Pillinger has only until the end of October to present his financial plan to ESA.

The launch of Mars Express is scheduled for 2003 and cannot be delayed, as it depends on planetary alignments. If Pillinger cannot obtain funding, ESA is likely to close the option on the feasibility studies of Mars Express, due to start in November.

A.A.

New faces, old policies in German coalition

[MUNICH] Edelgard Bulmahn of the Social Democrats (SPD) seems almost certain to become research minister in Germany's new SPD–Green Party coalition government (see *Nature* 395, 208; 1998). Wolf-Michael Catenhusen (SPD), a politician with long experience of science policy issues, will act as a state secretary with special responsibilities that include higher education and biotechnology.

Bulmahn has already said that one of her first acts will be to seek an increase in the research budget. But the winds of change are unlikely to blow very hard through the corridors of the ministry following last month's defeat of the Christian Democrats (CDU) after 16 years in power.

There is little to differentiate SPD and CDU research policies, apart from the SPD's greater emphasis on renewable energy and environmental research, and the promise — which most view with scepticism — of a significant increase in research funding.

Bulmahn is seen as both bright and personable. A rising star in the SPD, she was recently named the successor to Chancellor Gerhard Schröder as party leader in her home state of Lower Saxony.

Trained as an English teacher, she entered parliament in 1987 and immediately threw herself energetically into research politics. But for many years she stood in the shadow of Catenhusen, who chaired the all-party parliamentary committee for research and technology from 1987 to 1994 and is regarded as one of Germany's leading political experts on biotechnology.

Bulmahn was deputy chair of the research committee from 1990, before serving as chair

Research reactor escapes Green axe... so far

[MUNICH] Germany's Greens have failed to persuade the Social Democrat Party (SPD), as part of the coalition agreement between the two parties, to take a strongly critical stand on a research reactor being built at the Technical University of Munich. But the two parties did agree last week to phase out the use of nuclear power.

The reactor, FRMII, has come under fire in the past from both political parties for its planned use of highly enriched uranium (HEU) as a source of neutrons. Critics say the fuel could be diverted to make nuclear

weapons (see *Nature* 379, 284; 1996).

The reactor's operating licence is due to be issued by the Bavarian environment ministry next year. In theory, the federal government could veto Bavaria's decision.

"We are not worried about it," says Gert von Hassel, a spokesman for the Technical University of Munich. He says any attempt by the government to stop the reactor from operating would be strongly challenged.

Wolfgang Stöffler, SPD spokesperson on research policy, says the exclusion of

FRMII from the coalition treaty between the two parties does not automatically mean the new government has ruled out a possible veto.

No period has been set for the phasing out of nuclear power — responsible for 36 per cent of Germany's electricity. The costs of such a move could be extremely high. Germany's nuclear industry has said it would sue the government for hundreds of billions of Deutschmarks if it were to revoke the industry's unlimited licences to operate the country's 19 atomic plants. **Quirin Schiermeier**

in 1995 and 1996. In 1995 she became chair of the *Wissenschaftsforum der Sozialdemokratie*, a forum organized by the SPD to educate politicians in science issues. In 1996 she became SPD speaker on research affairs.

Like her predecessor as research minister, Jürgen Rüttgers, Bulmahn believes in the importance of competition, and says she will extend ideas he introduced, such as the *Bioregio* competition that pitted regions against each other in their bids to be recognized as most competent in biotechnology.

Also like Rüttgers, she is committed to research as a tool for fighting unemployment

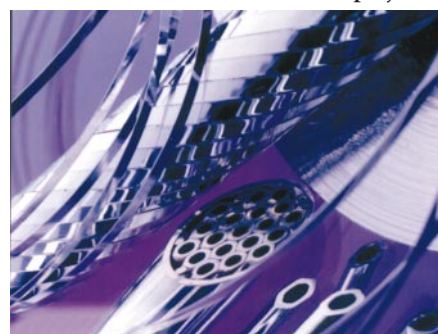
by promoting innovation. Yet she strongly supports basic research, and was one of its most vociferous defenders in the parliamentary research committee when it came under attack by industrialists in the early 1990s.

Bulmahn is generally well liked by colleagues from other parties. But her honeymoon period may be short, particularly as she plans to immediately put to the test the SPD's election promise to increase the research budget by asking for more than Rüttgers had proposed for the 1999 budget. And rifts with the Green Party are likely to emerge over biotechnology issues. **Alison Abbott**

Superconductors make their public debut in US utility network

[WASHINGTON] The US Department of Energy this week announced the first project to install high-temperature superconducting (HTS) power cables in a utility network. In two years' time, three 130-metre-long HTS cables will replace underground copper cables in an electricity substation in Detroit.

The \$5.5 million demonstration project



Bright future: superconducting components like these promise reduced transmission costs.

is jointly funded by the department and private industry. The participants hope that it will be a stepping stone to the large-scale commercialization of the high-temperature oxide superconductors first discovered 12 years ago (see *Nature* 386, 115–118; 1997).

The 24,000-volt multifilament cables will be manufactured by Pirelli Cables and Systems from HTS wire produced by the American Superconductor Corporation. They will provide three times the power-carrying capacity of the copper cables while occupying the same space.

The contract, and others in the energy department's Superconductivity Program for Electric Systems, constitutes a vote of confidence in HTS technology. The technology has followed a rocky road since widespread hopes of 'electricity too cheap to meter' were sparked by the discovery, in 1987, of the first superconductor to operate at liquid-nitrogen temperature.

These early hopes were soon dashed by the difficulties of making wire out of the brittle oxide materials, and by obstacles to achieving high current-carrying capacity. But ten years of work on the materials has led to recent successes, including laboratory demonstrations of 50-metre lengths of HTS multifilament cable by Pirelli and by two Japanese companies, Sumitomo Electric Industries and Furukawa Electric.

But the commercialization of HTS power cable still faces an important economic obstacle. Potential manufacturers have estimated that the cost of HTS wire (per ampère metre) would have to decrease 30-fold for power applications to become widespread.

The energy department's programme, which is funded at \$32.5 million for the fiscal year 1999, aims to help industry overcome this and other obstacles to the commercialization of HTS. **Laura Garwin**

Lobbyists elated as the NIH wins \$2bn budget increase

[WASHINGTON] US biomedical research is set to receive an unprecedented infusion of cash. President Bill Clinton and the Congress finally agreed last week to a budget for 1999 that will increase the funding of the National Institutes of Health (NIH) by \$2 billion, to \$15.6 billion.

This 15 per cent increase is the amount requested for NIH last December by the Federation of American Societies for Experimental Biology (FASEB), but which few of its members expected to attain. NIH institutes have been planning for months, however, in expectation of an increase greater than the 8 per cent that Clinton requested for the agency in February.

"We're elated over the NIH budget," says Bill Brinkley, president of FASEB and vice-president of Baylor College of Medicine at Houston, Texas. The money "can be spent well," he adds, predicting that NIH funding "is now on line for doubling in five years, from \$13 billion to \$27 billion".

The budget includes increases of around 15 per cent for most NIH institutes and centres. Of the larger ones, the National Heart, Lung and Blood Institute and the National Center for Research Resources do best, with rises of 17 per cent and 22 per cent respectively.

The final agreement is even more gener-

ous than the highest NIH budget brought to the negotiating table, the one proposed by the Senate. An extra \$40 million of research spending was released at the last minute, with \$15 million of it going to the National Human Genome Research Institute, whose budget will grow by more than 20 per cent.

When Clinton first proposed an 8 per cent increase for NIH, the extra funding was supposed to be contingent on legislation to introduce new tobacco taxes (see *Nature* **391**, 521–522; 1998). This legislation failed but, as expected, Clinton and the Congress managed to find a way to appropriate spending far in excess of the legally-binding budget caps they had agreed last year.

By designating \$20 billion as "emergency spending", the two sides were able to fund most of each others' priorities, while technically remaining inside the law.

Not every such priority was funded, however. Clinton's Climate Change Technology Initiative will develop more slowly than he had planned. This research programme is supposed to help the United States reduce its greenhouse-gas emissions, and has been opposed by the Congress on ideological grounds. Clinton requested an extra \$450 million for the initiative in 1999. Although details remain sketchy, he is expected to obtain about half that.

Colin Macilwain



Meldrum: still in the limelight.

Top BSE official denies charges of excessive secrecy

[LONDON] Keith Meldrum, the British government's chief veterinary officer during the crisis over bovine spongiform encephalopathy (BSE), last week defended his handling of the affair to a public inquiry.

One of the most pivotal figures during the crisis, Meldrum strenuously denied that the Ministry of Agriculture, Fisheries and Food put the interests of industry above those of public health, and refused to criticize the conduct of ministers.

"Those that would suggest that we tended to adopt policies that would favour the agriculture industry and to their advantage and to some extent therefore ignore the consumer would be totally wrong," said Meldrum. He did, however, acknowledge that, unlike the chief medical officer, he was unable to articulate a view that went against government policy.

Meldrum was speaking on Monday (19 October) at the start of his two-day appearance before the BSE inquiry in London, which was in its 68th day. Meldrum told the inquiry that, although he supported the principle of external advisory committees, he did not always agree with the need for them, particularly where relevant expertise existed within government laboratories.

Meldrum also claimed that the ministry had an "open policy", and "attempted as far as we were able" to keep all parties fully informed of all aspects of BSE. But he said that, although committed to giving the public "the full facts", he felt that the media did not always convey all the relevant details.

"How do you ensure that the consumer, the public, is fully informed about the risk and risk assessment to look at the safety of food?" he asked. "The problem is that there is no direct route to the public except through the media. You can produce leaflets, but it is difficult to get the public to read them."

Meldrum said he would reconsider his decision not to provide the committee with access to a complete tape recorded interview on the BSE affair he gave to researchers from the University of Surrey.

Ehsan Masood

Spain's cardiologists rally to Moncada

[BARCELONA] The Spanish Society of Cardiology (SEC) plans to make a formal protest about last week's decision by the Nobel committee to exclude Salvador Moncada from those awarded this year's prize for physiology or medicine.

The award was made to Robert Furchgott, Louis Ignarro and Ferid Murad for the discovery and elucidation of the biological functions of the gas nitric oxide (see *Nature* **395**, 625–626; 1998). No mention was made of Moncada, who has been a prominent figure in the study of the properties of nitric oxide since demonstrating in 1987 that it is identical to Furchgott's endothelium-derived relaxing factor.

Moncada was born in Honduras and educated at the University of El Salvador, and is currently at the Wolfson Institute for Biomedical Research at University College London. He has long enjoyed close links with cardiology researchers in Spain, partly as a result of a common linguistic and cultural background.

Members of SEC were preparing to

release a formal protest to the Nobel committee on Wednesday (21 October) to coincide with the opening of the National Congress of Cardiology in Malaga.

Alfonso Castro Beiras, president of the society and head of the Department of Cardiology at the Juan Canalejo Hospital at La Coruña, says that the note is intended to express SEC's unhappiness that Moncada's contributions to the field of nitric oxide research appear to have been ignored. According to Castro Beiras, many members of the society are expected to voice their criticism during the meeting.

Ironically, Enrique Castellon, deputy head of Spain's Ministry of Health, announced last week that the country's 1999 budget for public health will include an extra US\$2.6 million to create a centre for cardiovascular research. The centre will be along similar lines to the recently established National Institute of Oncology, headed by Mariano Barbacid. The new centre's main focus will be studies of nitric oxide as a signalling molecule in the cardiovascular system.

Xavier Bosch

Closer links urged for islands in the sun

David Dickson

The tenth anniversary of the Caribbean Academy of Sciences has highlighted both the case for greater regional collaboration on science-related issues, and the political and economic obstacles that can stand in the way.

[PORT OF SPAIN, TRINIDAD] If there was a silver lining to the clouds that raced across the Caribbean during last month's hurricane, it was perhaps that the damage the hurricane left in its wake underlined yet again the extent to which the islands of the region face common threats.

Researchers in the Caribbean are already working together to transform disaster mitigation efforts into risk management strategies. Others are exploring different forms of collaboration, from joint research on common marine problems to programmes to promote the popularization of science.

Such cooperation is being encouraged by foreign aid bodies, scientific organizations keen to provide assistance, such as the American Association for the Advancement of Science, and political groupings, such as the Organization of American States (OAS). They see the logic of increased regional cooperation, particularly in achieving economies of scale in tackling common scientific or environmental problems.

But the message has had a rougher reception at the national level. Not only are many Caribbean states yet to be convinced that investment in science and technology should be a political priority, but historical and linguistic rivalries remain close to the surface.

Both obstacles continue to be faced by a group of scientists who, ten years ago, launched a vigorous attempt to address them head-on through the creation of a Caribbean Academy of Sciences.

At the academy's tenth anniversary meeting in Port of Spain, Trinidad, last month, its president, Wilfred Chan, professor of chemistry at the University of West Indies in Port of Spain, told how the academy already has an impressive list of achievements to its name.

Over this period, the number of elected fellows has almost doubled to just under 200, researchers have been given support to go to international conferences, a programme has been developed to boost science education in schools, and a young-scientist award has been launched jointly with the Third World Academy of Sciences.

In an attempt to overcome some of the cultural differences between the islands in the region, what began as an English-speaking organization has made a strong effort to



Wind of change: collaboration in hurricane prediction points the way for joint projects.

broaden its turf. It now has members from several French-speaking countries — its last annual general meeting was held in Guadeloupe — and the annual meeting in two years time will be held in Havana at the invitation of the Cuban government.

Partly to acknowledge these achievements, last month's meeting was attended by representatives of counterpart bodies from around the world. Each expressed eagerness to support the academy's bid to play a greater role in promoting the development of science in the Caribbean region.

Christopher Thomas, assistant secretary-general of the OAS, said the academy's objectives aligned closely with his organization's view that "a collective approach by the region will provide the optimum thrust at this most critical moment of [its] development".

He said that Caribbean nations should pool their natural and human resources "in order to convert the region from a traditionally agricultural-based economy to one that is transformed through the application of science and technology".

Several science ministers also provided support. "The need to work together is unquestionable," said Rosa Elena Simeon, Cuba's minister of science, technology and

environment. As President Fidel Castro had made clear during recent visits to Jamaica, Barbados, Grenada and the Dominican Republic, she said, her country had already expressed its "will and decision" to share its human resources and knowledge with other Caribbean states.

Trevor Sudama, minister of planning and development for Trinidad and Tobago, said the academy's role "must be strengthened". But if the spirit has been willing, the follow-through has been relatively weak.

Ramsey Saunders, professor of physics at the University of West Indies and the academy's founding president, points out that, immediately before the academy was set up in 1988, the region's ministers for science and technology had jointly promised to support it. But this promise was never kept, and the academy continues to operate through the voluntary work of its members.

The caution of individual governments is not only economic. Officials point out that there are already a number of intergovernmental mechanisms for encouraging greater scientific collaboration in the region, in particular through the Caribbean Council for Science and Technology.

Proposals to set up an independent regional research council to agree common priorities and distribute research grants on a Caribbean-wide basis have therefore been received with little political enthusiasm.

Harold Ramkissoon, a former president of the academy and now its foreign secretary, supports the idea, arguing that such a research council could help to raise the scientific performance of the region. "We have a number of research organizations in the region and there is sometimes duplication of effort," he says. "A regional research council would be able both to avoid this and to mount regional research projects."

But others are less convinced. Government officials point to problems of political accountability, while some argue that it could detract from other goals — such as promoting the public understanding of science — that can still only be addressed nationally.

There is agreement, however, that the academy is helping to develop a regional strategy whose role can only grow in importance. This message is likely to be highlighted at next year's World Science Conference, being organized in Budapest by the United Nations Educational, Scientific and Cultural Organization (Unesco) — which supported the academy's anniversary celebrations — and the International Council for Scientific Unions (ICSU).

"With very limited funding, we have been able to mount a range of activities and establish links with other academics and international organizations world-wide," says Ramkissoon. "With regular funding, we could make a much greater contribution to both national and regional development." □

Dutch challenge to Brussels over biotech directive

[MUNICH] The Dutch government has appealed to the European Court of Justice against the European Union's new directive on the protection of biotechnological inventions, which allows patenting of human genes as well as transgenic plants and animals (see *Nature* 393, 200; 1998).

The Netherlands had voted against adoption of the directive following a resolution of its parliament two years ago condemning the directive on ethical grounds. The Dutch government is now arguing on technical grounds that the directive should have been subject to unanimous rather than majority voting.

But European Commission officials say the appeal is likely to be rejected. A similar appeal by Spain against a directive relating to legal protection of medicinal products was rejected by the court in 1995.

US panel to advise on women in science

[WASHINGTON] A national commission is to be established in the United States to advise the government on how women and

minorities can be encouraged to succeed in science and engineering, under a law signed last week by President Bill Clinton.

The commission will consist of seven representatives of industry and four academics, and will be asked to "identify and address the problems associated with the recruitment, retention and advancement" of women and minorities in science and engineering. It will be selected within 90 days and then has a year to prepare its report for the Congress and the president.

The law establishing the commission — which was suggested by the National Science Foundation in 1996 — was initially sponsored by Connie Morella (Republican, Maryland), chair of the technology subcommittee of the House of Representatives Science Committee. It was passed by the House last month and then by the Senate, where the measure was put forward by Senator Olympia Snowe (Republican, Maine).

French public say 'non' to modified organisms

[PARIS] A little over half of the respondents to a survey in France expressed negative feelings about genetically modified organisms, according to the opinion polls company TMO.

One quarter of the 1,000 individuals polled associated genetically modified organisms with danger, and one-fifth considered that they were "against nature". One quarter did not feel strongly either way, classifying the issue as a neutral scientific question.

Only 9 per cent of those surveyed expressed a positive opinion about genetically modified organisms.

Germany's PhDs will see their work go online

[MUNICH] Germany's main research grant agency, the Deutsche Forschungsgemeinschaft, is sponsoring research for a digital archive that would provide online access to research carried out by PhDs anywhere in Germany.

The first results, including instructions for authors, were presented at the Frankfurt book fair last week (for further information see www.educat.hu-berlin.de).

University libraries and five scientific societies, including the German Physical Society and the Society of German Chemists, are collaborating on the archive. They hope to develop software tools that would allow a directed search for components of online documents, such as specific terms, key words or references, that can be indexed.

NASA spacecraft to test new technology

[WASHINGTON] Deep Space 1, the first in a new line of US space missions designed to test advanced tools and techniques for scientific spacecraft, is scheduled to launch from Cape Canaveral, Florida on 25 October, and perhaps sooner. The mission's research objective, to fly past and photograph the asteroid 1992KD at a range of five to ten kilometres next July, is secondary to its testing of 12 new technologies.

These include a combination camera-spectrometer, autonomous navigation capability, and an ion propulsion engine that could provide cheap transport for future planetary spacecraft. Most of the technology evaluation for the \$152 million mission will take place in the two months after launch.

Sydney museum boss outlines research vision

[SYDNEY] Michael Archer, currently a professor of biology in the University of New South Wales, has been appointed director of the Australian Museum in Sydney. He aims to enhance the museum's reputation for research in zoology, marine science, anthropology and Earth sciences by

associating the museum with the preservation of the rich Riversleigh and Murgon sites in Queensland, through which he and his research students have revolutionized the dominant view of Australian pre-history.

Archer also intends to harness the museum's skills in biodiversity research by mounting a large-scale demonstration of the sustainability and economically competitive use of land through harvesting (not farming) native kangaroos, displacing the grazing of sheep, which has caused immense damage to Australia's fragile soils.

India bows to pressure on patents legislation

[NEW DELHI] India has signed the Paris Convention for the protection of intellectual property and has agreed to abide by the Patent Cooperation Treaty. The move ends 30 years of opposition based on the argument that it would destroy Indian industry. Both treaties will be binding on India from 7 December.

Ragunath Mashelkar, secretary for industrial research, says it will benefit domestic research and development, as it will now be possible to submit a single patent application valid in several countries (see *Nature* 394, 709; 1998). But critics who warn

of a sudden rise in drug prices say India has been trapped because it cannot get out of the treaty for at least five years.

Archimedes manuscript to go under the hammer

[LONDON] An original manuscript written by Archimedes in the third century BC, which was erased more than 1,000 years later and overwritten as a Greek religious text, is to be auctioned in New York next week. The 174-leaf manuscript contains the original text of Archimedes' treatises, including *On the Method of Mechanical Theorems* and *On Floating Bodies* (see below). The manuscript is expected to fetch between \$800,000 and \$1.2 million. It is known as a palimpsest, a

text written on vellum leaves from which the writing is washed away or scraped off so that the surface can be reused. Archimedes' writing was restored using digital scanning techniques.



Where has Britain's plutonium gone?

Sir—The UK government's Department of Trade and Industry has issued a document that claims to be "...part of the UK's renewed commitment to improve transparency and openness in its management of our national holding of civil plutonium"¹. In fact the new format provides considerably less information than previously. It no longer gives the reactor-specific data which enabled outsiders like ourselves to make independent assessments. And the total plutonium produced by the UK civil Magnox reactors is still not published.

Until recently, the UK government provided a subtotal of the civil Magnox plutonium. The unpublished balance was sent to the United States before 1971 under "mutual defence agreements" (MDA). Previous Conservative administrations in the United Kingdom maintained that this plutonium was not used in weapons, but refused to quantify the balance.

In 1985, we estimated that this balance was 6.3 ± 0.8 tonnes (refs 2, 3). We revised this figure to 5.4 ± 0.8 tonnes (ref. 4) when it was revealed that the total plutonium in solid waste was much larger than official sources had indicated previously.

In February 1996, the US Department of Energy published an inventory of its plutonium stocks, stating that "... from 1959 to 1980, the US acquired a total of 5.4 tonnes of plutonium in exchange for 6.7 kg of tritium and 7.5 tonnes of highly enriched uranium"^{5,6}. The total of 5.4 tonnes is in remarkable agreement with our revised

figure published four years earlier.

However, refs 5 and 6 did not specify the end uses of the plutonium in the United States, nor did they clarify how much originated in the UK civil Magnox reactors rather than the military ones at Calder Hall and Chapel Cross.

The question of the UK origin of the plutonium is problematic because civil and military Magnox spent fuel was 'co-processed' at Sellafield until 1986. In December 1997, a US press release⁷ was issued stating that 0.1 tonne of very high purity plutonium (2% plutonium-240 content) was sent to the United States under the exchange programme and that no other plutonium with plutonium-240 content less than 10% was involved. This suggests that the amount of military-origin plutonium involved in the exchanges was small, consistent with estimates that the total UK inventory of weapons-grade plutonium is relatively small. (Calder Hall and Chapel Cross have not operated on a military cycle for many years.)

We believe that the amount of plutonium of military origin sent to the United States under the MDA was a small part of the 5.4 tonnes, possibly as little as the 0.1 tonne of 2% plutonium-240 purity referred to in ref. 7. We conclude that the United Kingdom provided the United States with around 5.4 tonnes of plutonium from the UK civil stockpile.

All this 5.4 tonnes plutonium is in the US defence stockpile. The UK government maintains it has not been used in weapons,

but the US civil destinations that have now been listed contain at most 4 tonnes of plutonium^{2,3}, leaving around 1.4 tonnes not accounted for. If sufficiently pure, this amount of plutonium could provide up to 300 warheads.

We call on the present (Labour) UK government to clarify fully the past history. The US disclosures have removed the previous (Conservative) UK government's excuse for secrecy, which was because it would reveal the quantity of highly enriched uranium received by the United Kingdom for defence purposes. The United States published the exact figure for the highly enriched uranium involved in the exchanges two years ago, and it does not appear that the security of the realm has been imperilled.

K. W. J. Barnham, J. Nelson

Physics Department, Imperial College of Science, Technology and Medicine, London SW7 2BZ, UK

R. A. Stevens

Join Systems, Menlo Park, California 94025, USA

1. "Annual plutonium figures for UK" (UK Department of Trade and Industry press release P/98/424, 2 June 1998).
2. Barnham, K. W. J., Hart, D., Nelson, J. & Stevens, R. A. *Nature* **317**, 213–217 (1985).
3. Barnham, K. W. J., Hart, D., Nelson, J. & Stevens, R. A. *Nature* **333**, 709–710 (1988).
4. Barnham, K. in *Plutonium and Security* (ed. Barnaby, F.) (Macmillan, London, 1992).
5. *Plutonium: The First 50 Years* (US Department of Energy, February 1996).
6. "Openness, press conference fact sheets" (US Department of Energy press release, 6 February 1996).
7. US Department of Energy press release R-97-141 (22 December 1997).

Journals are best left on the shelf

Sir—I read with interest the announcement that *Nature* would be offering an electronic version of the journal free to subscribers (*Nature* **395**, 415; 1998). I very much enjoy the journal and expect the electronic version to quickly gain a wide readership.

However, the issue of electronic publishing is still contentious. It seems to boil down to this: publishers want to encourage use of electronic versions but still maintain cost-effectiveness, while subscribers (especially institutional subscribers) enjoy such benefits as rapid searching, but want guarantees of permanence. I have no doubt that articles published electronically in journals of high standing will be reasonably 'permanent', in that someone (preferably

the publisher) is likely to always maintain a copy.

The discussion of electronic permanence has missed one important aspect: that of permanent availability. If a library subscribes to the electronic version of a journal for, say, five years, it gains the use of the text for those five years. But if budget cuts force the subscription to be dropped, the library's patrons no longer have access to the five years purchased, whereas a library that had ordered a print subscription would always have at least those years paid for on the shelf.

The bottom line argument seems to be the willingness of the scientific community and publishers to make the transition from journals as an item, to journals as a service. I am not willing to invest the cost of several years of a journal subscription in a service that will disappear as soon as I stop paying the fees. A paper journal, on the other hand, can remain on my shelf indefinitely,

independent of changing financial situations.

Anthony K. Grafton

Campbellsville University, Campbellsville, Kentucky 42718, USA

Hybrid journals

Sir—Science and research libraries in India continue to spend a large proportion of their budgets on subscriptions to printed versions of journals, but frequently cannot subscribe to electronic-only or electronic versions of print journals owing to lack of infrastructure facilities and access to the Internet. Discussions on this topic have overlooked another important development in journal publishing — what is termed the 'hybrid' journal.

In a hybrid journal, the main information is delivered in print whereas

additional or supporting material is available in print and electronically. In a recent development, some journals are providing this supporting material in electronic form only (several journals of the American Chemical Society, for example), and this is a cause of concern for research workers in India and other developing countries.

Because many libraries in India still do not have Internet access, and many researchers do not have any other form of organized Internet access, subscribers to a hybrid journal may have only partial access to its contents. We checked the *National Union Catalogue of Scientific Serials in India*, and find that nearly 40% of Indian libraries that subscribe to some hybrid journals do not have access to electronic information.

We would like to make the following proposals to tackle this problem. First, information-service providing agencies, such as INSDOC or the National Centre for Science Information, could identify hybrid journals with Indian subscribers and establish formal mechanisms for procuring and making available such supporting material in a convenient format and medium on a regular basis. Alternatively, libraries without these facilities could link up with libraries with these facilities for the supply of such materials so that both the libraries and science workers can obtain full access to what they have already paid for. Such services by these agencies could also in principle cover electronic-only journals (subject to copyright and other regulations), so that the gap between the 'haves' and the 'have-nots' is reduced.

Second, we suggest to the publishers of hybrid journals that the type of information available in supporting material is clearly indicated in the printed version so that the reader can decide whether to request it. Not all journals clearly specify the nature of the supporting material.

Finally, secondary sources such as *Chemical Abstracts* and *Current Contents* are acknowledged means of identifying relevant material, but generally do not indicate the existence of supporting material. Hence, the availability of such material is known only when the original article or its photocopy/reprint is seen. It would be useful if secondary sources indicate the existence of such material.

B. Annapoorani, S. Amba

Information Area,
Central Leather Research Institute,
Chennai 600 020, India

K. S. Raghavan

Department of Library and Information Science,
University of Madras,
Chennai 600 005, India

Popular reaction against science

Sir — The editorial on *The X Files* made some excellent points — especially about the rigour of Mulder and Scully's investigative methods (*Nature*, 394, 815; 1998). The editorial went on: "The popularity of *The X Files* suggests that the public clearly has more of a feeling for the spirit of scientific enquiry than some give it credit for".

Although I would like to believe this statement, I think it is false. The success of *The X Files* is part of a popular reaction against science. Many cultures (particularly in the West) are increasingly secular, and there is a prevailing feeling that there is no longer any mystery that science cannot elucidate. *The X Files* offers the comforting spectacle of science *not* coming up with answers — in fact, of frequently falling flat on its face. The programme's massive popularity results much more from the decline of the church (formerly a wellspring of mysticism), and scientific rumours of an imminent 'Theory of Everything', than from a true spirit of scientific enquiry.

Daniel Davies

The Lancet,
42 Bedford Square,
London WC1B 3SL, UK

Two-way street

Sir — Students of phylogeny cannot but welcome the application of their methods to the discovery of relationships between families of manuscripts, as in the recent study of 58 fifteenth-century manuscripts of "The Wife of Bath's Prologue" from *The Canterbury Tales*¹. It may prove of interest to philologists and phylogeneticists alike to realize that progress may also be made in the other direction.

For example, in their analysis of protistan phylogeny, Ragan and Lee² introduced to the biological literature a parsimony-based procedure originally developed³ to detect textual contamination, a philological equivalent of the lateral transfer of genes (or characters).

Historians wishing to search for the history of current phylogenetic approaches may profitably look into the works of Renaissance scholars such as Angelo Poliziano (1454–94). In his textual reconstruction of classic Greek and Latin works, Poliziano⁴ discounted the then fashionable method of relying on the textual version preserved by most extant manuscripts. He advocated instead careful

comparative weighting of evidence, because close replicas of the same version or other interdependent sources should not be given the same individual weight as largely independent sources. This is one of the most fundamental principles of current phylogeny-based approaches to comparative biology⁵.

Alessandro Minelli

Department of Biology, University of Padova,
Via Ugo Bassi 58B, I 35131 Padova, Italy
e-mail: almin@civ.bio.unipd.it

1. Barbrook, A. C., Howe, C. J., Blake, N. & Robinson P. *Nature* 394, 839 (1998).
2. Ragan, M. A. & Lee, A. R. in *The Unity of Evolutionary Biology* (ed. Dudley, E. C.) 432–441 (Dioscorides, Portland, 1991).
3. Lee, A. R. *Studia Patristica* 20, 24–32 (1989); in *Computers in Literary and Linguistic Research* Vol. 3 (ed. Choueka, Y.) 261–292 (Champion-Slatkine, Paris, 1990).
4. Grafton, A. *Defenders of the Text* (Harvard Univ. Press, Cambridge, MA, 1991).
5. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology* (Oxford Univ. Press, 1991).

Referencing crystal coordinates

Sir — Analyses of the crystal structures of macromolecules, the coordinates of which are deposited in databases, constitute a growing component of the biochemical literature. Given the magnitude of investment that has been made in determining protein and nucleic-acid structures, this is entirely appropriate; the gold hidden in that vast mine of information needs to be recovered. But I am disturbed by a practice of the community that engages in coordinate analysis: its use of database identification numbers as the sole reference to the coordinate sets discussed in publications.

This is not good scholarly practice. It conceals from readers the identities of the scientists responsible for the coordinate sets used, and makes it difficult for readers to find primary references. It is also unfair. For better or worse, citation indices are increasingly used to evaluate the contributions scientists have made to their fields. No credit will accrue to those who made the effort to determine a structure unless the papers that make use of its coordinates include a proper reference.

For these reasons, the *Biophysical Journal* will from now on require authors to include in their papers full references for all the coordinate data sets they have used, as well as database identification numbers. I hope that other journals will institute similar policies to keep what is now a modest problem from getting out of hand.

Peter B. Moore

(Editor, *Biophysical Journal*)
9650 Rockville Pike, Bethesda,
Maryland 20814, USA

Adapting to the inevitable

Greenhouse-gas emissions targets to be discussed in Buenos Aires next month will have little effect on the potential impacts of climate change. We should be exploring ways of adapting to impacts, some of which are inevitable.

**Martin Parry, Nigel Arnell,
Mike Hulme, Robert Nicholls
and Matthew Livermore**

In Kyoto last December, at the third conference of the United Nations Framework Convention on Climate Change, targets were agreed for reductions in greenhouse-gas emissions. On 2 November in Buenos Aires, negotiators will reconvene at the framework convention's fourth conference to agree the mechanisms and a timetable for implementation. We shall be hearing a good deal about trading permits, compliance and enforcement in the weeks to come. But, in reality, the control of global warming that can be achieved on the current agenda is very limited.

The Kyoto Protocol is an agreement to a 5.2 per cent reduction in greenhouse-gas emissions by about 2010 (relative to 1990), and constant emissions thereafter. But these targets only relate to the so-called annex 1 countries (38 industrialized nations). These countries together account for about 57 per cent of present global carbon emissions, but will produce only 25 per cent of the emissions growth over the next 20 years. Most future growth in emissions is expected to occur in the fast-developing regions of Asia and Latin America, which are not signatories to the framework convention.

As a consequence, the Kyoto target itself does relatively little to combat the rate of climate change. The warming expected by 2050, without any deliberate mitigation, is estimated by the Intergovernmental Panel on Climate Change (IPCC) at 1.4 °C with respect to the 1961–90 average¹. About 0.25 °C of this has already been realized by the 1990s. Our model predictions suggest that fully implemented Kyoto targets would reduce this global warming by 2050 only by about 0.05 °C. Even more radical targets, such as a 20 per cent reduction in greenhouse-gas emissions from annex 1 countries, would reduce it by only a further 0.1 °C by 2050.

These minor reductions in the expected warming mean that the projected impacts of change are barely affected. The global number of people put at increased risk of hunger, water shortage or coastal flooding during storms as a result of projected climate changes is hardly touched by the targets under discussion at Buenos Aires, even if full implementation of the targets is agreed there. The numbers in Table 1 are derived from models reported at Kyoto². Although, for example, an extra 23 million people could be affected by coastal storm flooding due to sea-level rise without any mitigation,

Table 1 Impacts estimated for the year 2050

Emissions scenario	Global warming (°C) with respect to 1961–90	People (millions) at additional risk of:		
		water shortage	coastal flooding due to global warming	hunger
Unmitigated	1.39	1053	23	22
Kyoto	1.33	1053	22	20
20% reduction	1.22	909	21	17
30% reduction	1.19	891	20	16

perhaps one million might avoid such flooding as a result of achieving the Kyoto target.

The convention calls on signatories to take action to safeguard food security, ecosystems and sustainable development from dangerous levels of climate change. But the current target does not do this. This does not mean that we should despair, but it emphasizes two things. First, Kyoto and Buenos Aires are only the first steps in a process that must involve much greater reduction in emissions and also, crucially, the participation of developing countries. In this respect, the achievements of the industrialized countries at Kyoto, if ratified, are important in providing a lead that will encourage others to follow. Second, mitigation by reducing greenhouse-gas emissions cannot be the entire response to the threat of climate change. Given the long history of past emissions from industrialized countries and the inertia of the climate system, we will experience a substantial amount of further climate change even if we make huge cuts.

We should, therefore, be thinking seriously about how we can best adapt to climate change. Although article 3 of the convention calls upon signatories to take steps to reduce climate impacts, adaptation has received very little attention compared with mitigation. This may be partly because to admit the need to adapt sounds defeatist to negotiators, and also because adaptation seems more complicated than mitigation (emissions sources are relatively few, but the array of adaptations is vast). Yet to ignore adaptation is both unrealistic and perilous.

Moreover, there are two very sound reasons why we should seek global agreement on adaptation. First, our current vulnerabili-

ty to existing climatic variability is very costly. For instance, about 640 million people are at risk of hunger now. Poverty is the root cause, but much of the year-to-year variability in hunger is due to drought. By drought-proofing those at risk now we could secure their present livelihood and reduce the impact of future climate change. There are many kinds of such 'win-win' solutions that serve both our present and future needs, such as increasing irrigation efficiency, breeding more drought-resistant crops and developing buffer stocks of food.

Second, adapting to climatic variability has a substantially greater effect of reducing impact than does mitigation. Consider, for example, the effect of reducing water demand, shown in Table 2 as being reduced in each country by 5 and 10 per cent below current projections for 2050. Reducing water demand by just 5 per cent has four times as great an effect as reducing emissions by 30 per cent. Broadly, the same stress-reducing outcomes would stem from similar demand reductions in other impact sectors (such as reducing soil erosion, or reducing crop yield losses to pests and diseases).

There is a risk that negotiators have lost sight of the ultimate objective of the convention, which is to avoid dangerous levels of climate change. Current mitigation targets will not achieve this and should not be mistaken for effective climate management. The other 'half' of the convention — action to reduce impacts — needs to be considered at the same time. □

Martin Parry and Matthew Livermore are at the Jackson Environment Institute, University College London, 5 Gower Street, London WC1E 6HA, UK. Nigel Arnell is in the Department of Geography, University of Southampton, Southampton SO17 1BJ, UK. Mike Hulme is at the Climatic Research Unit, University of East Anglia, Norwich NR4 7TJ, UK. Robert Nicholls is at the Flood Hazard Research Centre, Middlesex University, Enfield EN3 4SF, UK. They are lead authors of the IPCC.

1. Intergovernmental Panel on Climate Change. *Climate Change 1995: The Science of Climate Change* (Cambridge Univ. Press, Cambridge, 1996).

2. Department of the Environment, Transport and the Regions. *Climate Change and its Impacts* (DETR, London, 1997).

Table 2 Effects of reducing demand for water on global impacts of climate change by 2050

Emissions scenario	People (millions) at additional risk of water shortage following change in demand for water (%)		
	0%	–5%	–10%
Unmitigated	1053	445	131
Kyoto	1053	445	131
20% reduction	909	349	125
30% reduction	891	349	–200

A channelled plume under Africa

Geoff Davies

Large areas of east Africa are and have been highly active geologically, but the underlying processes are debated. New modelling work may shed light on those processes and also illustrates the growing maturity of such studies.

For Earth scientists, the African continent and its surroundings are a distinctive and rich laboratory. Among many noteworthy features, Africa has the highest mean elevation of all of the continents; it hosts the great east African rift system; there is young and active seafloor spreading in the Red Sea and Gulf of Aden; great volumes of volcanic deposits are found in the northeast; and minor volcanism and associated rifting of the crust are scattered across a huge area of the eastern, central and northern parts of the continent. Because all of these features are suggestive of warm or upwelling mantle under Africa, they have been seen for some time as potentially involving mantle plumes —

narrow, hot upwellings inside the Earth that were first proposed by W. Jason Morgan¹ in 1971.

Opinion has been divided as to how many plumes there might be under Africa. At one extreme, in 1976 Burke and Wilson² proposed up to about 40 volcanic 'hotspots' in Africa, but dynamicists debate whether so many plumes could rise beneath one continent. Most other counts were lower, with Duncan and Richards³ in 1991 including only one clearly identified hotspot within the African continent. However, if that were the case, it is not clear how all of the scattered volcanism could be accounted for by only one or a few plumes. On page 788 of this issue⁴, Ebinger and Sleep pro-

pose a way in which much of the volcanism might be explained by a single, large plume. The novel feature of their story is the way topography on the base of the lithosphere, some 100–150 km below the surface, might channel the plume material into streams and pools.

Mantle plumes are now widely thought to rise from a hot thermal boundary layer at the base of the mantle (Fig. 1). When a new plume starts, it is led by a large spherical 'head' as much as 1,000 km in diameter. As the head nears the top of the mantle, it pancakes under the lithosphere, which is the cool, strong outer part of the Earth comprising the crust and the upper part of the mantle down to 100 or 200 km depth. The pancaking plume head spreads to a diameter of about 2,000 km and thins to about 200 km vertically. For most of its ascent, the plume material is in the solid state, but its high temperature allows it to deform slowly and so to behave like a fluid on geological timescales. However, as this hotter material reaches shallow depths, it may begin to melt because of the reduction in pressure, producing magma that rises and erupts on the Earth's surface. It is widely believed that the arrival of a new plume head is the cause of 'flood basalt' eruptions that spew millions of cubic kilometres of magma onto the Earth's surface within the relatively short period of a million years. Worldwide, about a dozen of these giant eruptions are known to have occurred within the past 250 million years.

One of the younger flood basalts, about 30 million years old, is on the Ethiopian plateau, near the junction of the Red Sea, the Gulf of Aden and the east African rift system. The Afar hotspot (see map on page 789), inferred to be the trace of the thin plume 'tail' that followed the plume head (Fig. 1), is inferred to be still active in this region. The arrival of the Afar plume may not only have produced the Ethiopian flood basalts but also possibly triggered or promoted the rifting that radiates from the region.

Ebinger and Sleep's proposal is a bold extension of this picture. They suggest that a much broader region has been affected by the Afar plume, a region extending down the east African rift, offshore and south to the Comoros Islands near Madagascar, west to the Darfur uplift, and possibly much further to the Adamawa plateau and the Atlantic coast (see Figs 1 and 2 on pages 789 and 790). They explain the scattered nature of the associated volcanism as being due to variations in the thickness of the lithosphere through this region.

There are two aspects to their mechanism. First, the plume material will tend to rise into regions of thinner lithosphere because of its buoyancy (Fig. 1). Second, melting occurs once the plume material reaches a critical depth and where it is still

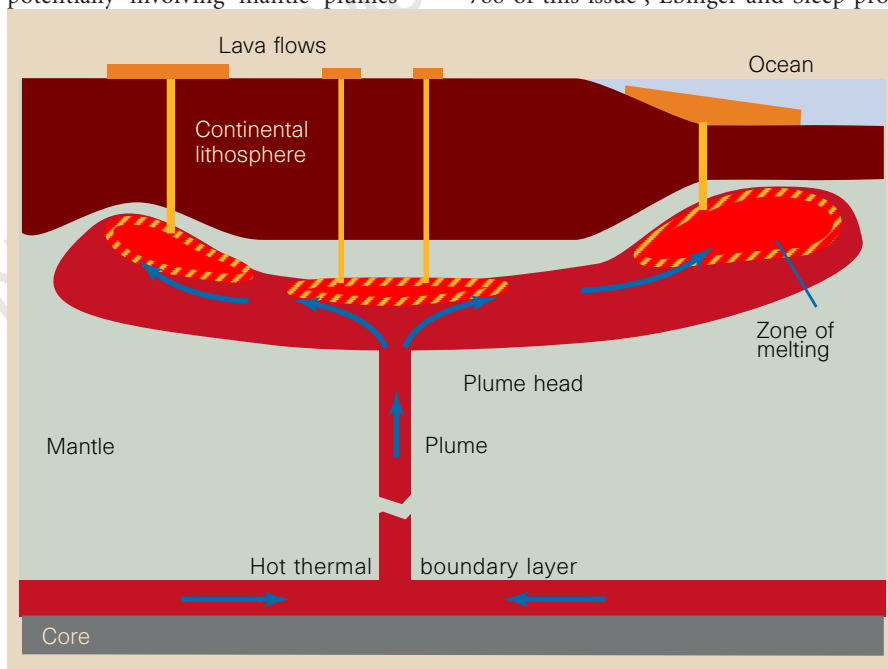


Figure 1 Ebinger and Sleep's⁴ proposal for how a mantle plume interacting with topography on the base of the lithosphere might explain scattered volcanism across a very broad region of east Africa. A large plume head followed by a thin plume tail rises from the base of the mantle. The head flattens and spreads as it approaches the base of the lithosphere, and it tends to flow up and under the thinnest parts of the lithosphere. Melting of the hot plume material only begins above a critical depth, and it only occurs as the plume material is rising and decompressing. Thus, melting tends to be concentrated where the plume material flows up a slope towards thinner lithosphere, as well as over the plume tail.

rising, so that the pressure it is under is decreasing. The latter point is crucial in determining where the melting will actually occur. It will not necessarily be focused where the lithosphere is thinnest, but rather where the base of the lithosphere slopes upwards. If the plume initially rises under thick lithosphere, it may melt very little or not at all (Fig. 1). If the material then flows horizontally, it will not melt any more. If later it reaches thinner lithosphere and can flow upwards again, it can melt further (Fig. 1). Subsequent rifting may trigger more melting, and may also promote the flow of plume material over greater distances.

Ebinger and Sleep's proposal is a stimulating one that is well worth exploring, but many uncertainties remain. Their modelling of plume flow and melting is plausible, but (appropriately at this stage) quite simplified. Geophysical information on lithosphere thickness is rather sparse. Subsurface magma, intruded into thick sediments, may be more extensive than is known at present, but more seismic data are required to find out. The details of melting relationships are not well constrained. Finally, the ages and geochemistry of the erupted rocks have some broad consistency with the idea, but much more detail needs to be filled in.

A potential complication is that George *et al.*⁵, in a paper published earlier this month, argue on the basis of new dating results that there are actually two plumes in this region. There were eruptions on the southern Ethiopian plateau as early as 45 million years ago which George *et al.* attribute to an older plume now located under Lake Victoria, which lies to the south. They ascribe later eruptions in southern Ethiopia, 19–12 million years ago, to the spreading influence of the more northerly Afar plume.

If George and colleagues' interpretation is confirmed, then Ebinger and Sleep's one-plume model would be a bit too simple, but their channelling mechanism still shows us a way in which scattered and apparently unrelated activity may have a common cause. In the meantime, Ebinger and Sleep's paper is a good illustration of the growing maturity of mantle dynamics modelling. The proposal is quantified, and testable in many respects. Its full evaluation will require a conversation between dynamists, geologists, geochemists and geophysicists. In other words, mantle dynamics is becoming a useful part of geology, in the broad sense. □

Geoff Davies is in the Research School of Earth Sciences, Australian National University, Canberra, ACT 0200, Australia.

e-mail: Geoff.Davies@anu.edu.au

1. Morgan, W. J. *Nature* **230**, 42–43 (1971).
2. Burke, K. C. & Wilson, J. T. *Sci. Am.* **235**, 46–57 (1976).
3. Duncan, R. A. & Richards, M. A. *Rev. Geophys.* **29**, 31–50 (1991).
4. Ebinger, C. J. & Sleep, N. H. *Nature* **395**, 788–791 (1998).
5. George, R., Rogers, N. & Kelley, S. *Geology* **26**, 923–936 (1998).

Ecology

Stability is woven by complex webs

Gary A. Polis

Ecological systems are extraordinarily complex. Each of the estimated 10–50 million species on Earth is directly affected by many factors — physical and biological, genetic and environmental, strong and weak, positive and negative, local and distant, current and historical. Add to this the fact that single factors operate jointly in nature, and the number of possible indirect and interactive effects on each species is astronomical. Dealing with such complexity is a central problem in ecology, full of the same contentious issues and rhetoric about reductionistic versus holistic approaches as are debated in other sciences. But on page 794 of this issue, McCann and colleagues¹ offer a new and valuable approach to understanding how complexity affects ecological systems.

In 1973, Robert May² unequivocally showed that, according to his seminal

models, 'complexity' is not stable — it will eventually cause an ecological system to 'fall apart', with strong fluctuations in abundance and the eventual loss of species. However, it was clear to empiricists and some theoreticians that natural systems are quite complex. In any one system, a great diversity of species is connected through many different interactions.

The idea that complex systems are somehow unstable held sway in the ecological literature for nearly two decades. A group of modellers used May's approach to reiterate that complexity in food webs — many connections between species, omnivory and long feeding chains, for example — was destabilizing. These theorists then collected statistics from published food webs as 'empirical support', which were thought to show that natural systems are quite simple at their cores. This claim agreed with their the-

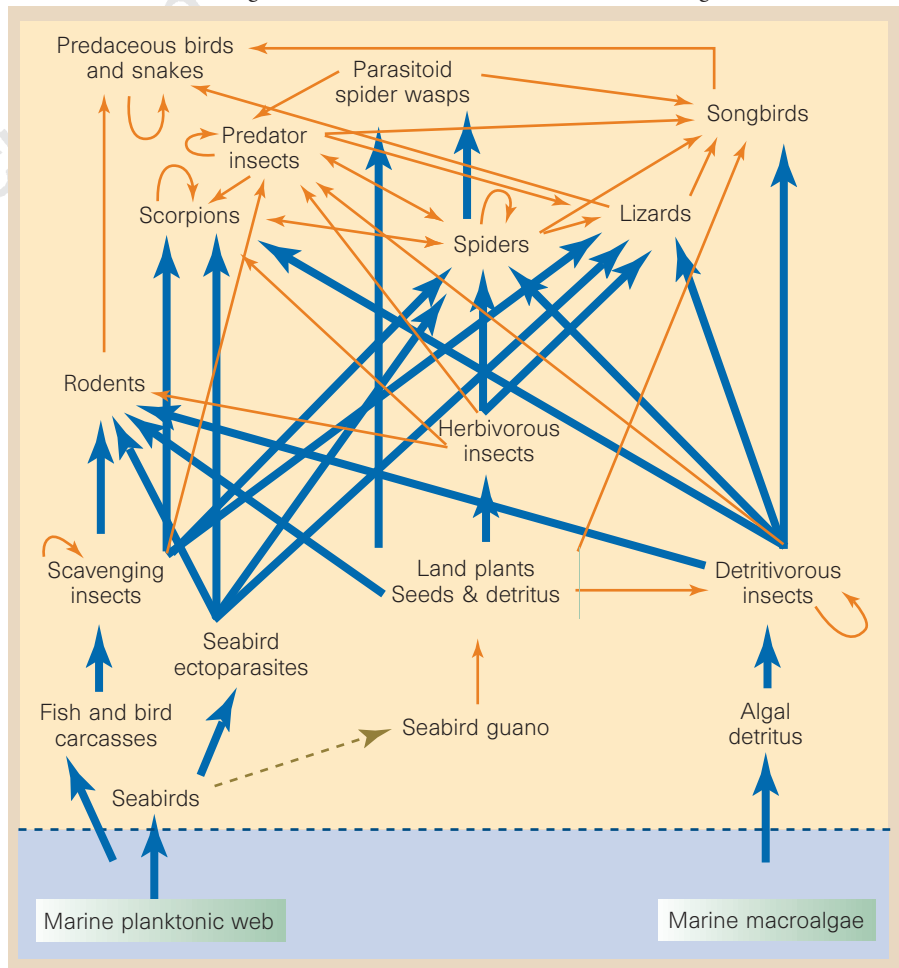


Figure 1 Food-web complexity, here for islands in the Gulf of California. McCann *et al.*¹ have shown that complexity in food webs facilitates stability in natural communities, a finding that runs counter to previous theory. This web depicts a subset of interactions on the islands in wet El Niño years^{12,13}; marine sources and land plants provide the fuel for an estimated 500–1,000 species. Thick lines represent strong interactions; thin lines, weak interactions.

ory that natural systems are kept so simple because more complex configurations of species reduce a community's ability to recover from perturbations. Thus, complex systems tend to decay or disappear.

Although the theory that nature abhors complexity ran counter to the intuition and understanding of most ecologists, there was little coherent evidence to argue against it. But this theory and its empirical evidence have since come under attack^{3–6}. It is now clear that real communities are brimming with complexity, and that the theory failed to describe reality. But although the more recent literature^{7–9} indicates that complexity provides an interwoven matrix that holds the rich tapestry of a community together, mathematical models showing the cohesive role of complexity have been few and far between^{10,11}. This is where McCann *et al.*¹ come in. They provide the mathematical and theoretical basis for the paradigm shift that is now overtaking ecology.

Using a set of innovative models, McCann and colleagues show that complexity tends to stabilize a system — that is, it dampens population fluctuations and prevents the loss of species. They find that the many weak links that connect a species to its community almost always stabilize the dynamics of that species. The authors derive this crucial finding repeatedly, using a variety of 'food-web modules' that represent different forms of common, complex interactions with direct and indirect effects. Examples of these modules include omnivory, consumers with many prey or with resources from a second habitat, or intra-guild predation (where the competitors are both predators and prey). All in all, their results are quite resonant with the view of most ecologists.

How did McCann *et al.* achieve these intuitively satisfying findings, so contrary to the theoretical dogma? Their models have a different structure and incorporate much more biologically reasonable assumptions than previous ones. First, unlike traditional Lotka–Volterra-based models, they incorporate nonlinear terms for per capita growth and consumption rates. Second, they do not assume that this system is at equilibrium (it is clear that processes and species abundances in most ecological systems are just as likely to be in flux as to remain constant). This means that the analyses do not use the traditional approach to search for stable equilibria, but instead focus on 'attractors' — processes that bound population oscillations away from low densities, thus making extinction much less likely. Third, their nonlinear terms incorporate well-substantiated behaviour for feeding interactions, such that consumers (predators and herbivores, for example) cannot maintain high feeding efficiency on many prey at the same time. This realistic constraint means that consumers will ease

up on a resource when it is at low densities. So, having many prey tends to prevent consumers from driving any particular prey to extinction — another big stabilizing effect.

Ecology is a young but maturing discipline. The findings of McCann *et al.* now put both theoreticians and empiricists 'on the same page' in their attempts to find out how natural systems operate. The understanding that complexity is vital to the integrity and stability of natural systems allows ecologists to argue, more coherently, why we must preserve the diverse elements and species that coexist in a healthy, sustainable and well-functioning ecological community. Indeed, as we enter what E. O. Wilson calls the "century of the environment", one crucial function of ecology is to provide an unbiased, scientific basis on which political and social decisions can be made about how best to treat our natural environment. This is a

daunting task, but ecologists are progressing very well. □

Gary A. Polis is in the Department of Environmental Sciences and Policy, University of California, Davis, California 95616, USA.
e-mail: gapolis@ucdavis.edu

1. McCann, K., Hastings, A. & Huxel, G. R. *Nature* **395**, 794–798 (1998).
2. May, R. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, 1973).
3. Cohen, J. E. & Newman, C. M. *J. Theor. Biol.* **113**, 153–156 (1985).
4. Winemiller, K. O. *Ecol. Monogr.* **60**, 331–367 (1990).
5. Polis, G. A. *Am. Nat.* **138**, 123–155 (1991).
6. Polis, G. A. *Aust. J. Ecol.* **19**, 121–136 (1994).
7. Strauss, S. Y. *Trends Ecol. Evol.* **6**, 206–210 (1991).
8. Polis, G. A. & Strong, D. R. *Am. Nat.* **147**, 813–842 (1996).
9. Fagan, W. F. *Am. Nat.* **150**, 554–567 (1997).
10. Huxel, G. R. & McCann, K. *Am. Nat.* **152**, 460–469 (1998).
11. McCann, K. & Hastings, A. *Proc. R. Soc. Lond. B* **264**, 1249–1254 (1997).
12. Anderson, W. B. & Polis, G. A. *Proc. Natl Acad. Sci. USA* (in the press).
13. Polis, G. A., Hurd, S. D., Jackson, C. T. & Sanchez-Piñero, F. *Ecology* **78**, 1884–1897 (1997).

Macromolecular chemistry

Limits to growth

Philip Ball

"Chemistry — What For?" The title of this recent conference* reveals something about the discipline. It is less likely that anyone would ask what biology, physics or Earth science is for — they are simply there, begging explanation. Chemistry is different. As Berthelot said, it "creates its own object" — it has the perpetual privilege of inventing itself anew. The conference's subtitle, "A View For the Next Century", announced an intention to explore what form the third-millennial incarnation will take.

But prediction is a risky business — especially, as Yogi Berra said, when it is about the future — and forecasts were muted. If there was any shared opinion, it was that the future of chemistry is about greater control over matter: not so much in the manner of a dictator's domination of his subjects, but of an artist's facility for determining what appears on the canvas.

One way to achieve this is by random mixing: the automated, computer-controlled version of traditional make'n'bake called combinatorial chemistry (D.M. Golden, SRI International, Menlo Park), which now provides not only candidate drugs but also phosphors, dielectrics, heterogeneous catalysts, superconductors and other solid-state materials. As there are not enough atoms in the Universe to make even a single copy of every possible 50-atom molecule, some rational planning will be needed to select the combinations. And for new materials, the need to optimize several sometimes conflicting parameters will place

heavy demands on screening methods.

More satisfying to the chemical purist would be atom-by-atom control of matter. Well-defined inorganic clusters can already be synthesized from the molecular to the micrometre scale: there are several methods of making nanocrystals of metals and semiconductors such as cadmium sulphide, for example. This technology can improve mechanical properties such as hardness, or exploit quantum-confinement effects to engineer absorption spectra and other optoelectronic properties.

But it is more difficult to control size and morphology down to the last atom — in other words, to make identical molecules (inorganic macromolecules, one might say) instead of more-or-less monodisperse little chunks of solid. Even so, this can be achieved, for example, with semiconducting copper chalcogenides Cu₂X, where X is sulphur, selenium or tellurium (H.-D. Fenske, Univ. Karlsruhe). The trick is to use precursor compounds containing organic ligands (such as P(C₆H₅)₃) that coat the surface of the cluster and inhibit nucleation of the bulk solid. Clusters up to Cu₄₂₈Se₂₁₄ have been prepared in this way.

These clusters can give us a partial answer to a fundamental question: when does nanoscale matter cease to be molecular and start acting instead like a piece of bulk material? Structure and properties do not necessarily evolve smoothly with increasing cluster size, as the bulk atomic structure might not be the most stable one at very small scales. For Cu₂Se clusters, a clear transition can be seen in passing from the Cu₆₉ species, which is sphere-like with a roughly

*Chemistry—What For? A View for the Next Century, Bielefeld, Germany, 30 September–2 October 1998.

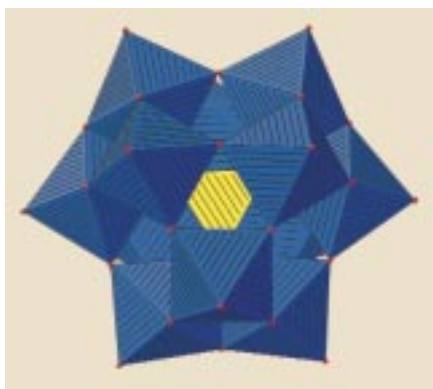


Figure 1 The ion $M_{12}O_{40}^{4-}$, a cluster of MO_6 octahedra linked via the oxygen atoms at the corners. In the centre of the cluster is a tetrahedrally coordinated site (light grey) where, for example, a silicon atom sits in $SiW_{12}O_{40}^{4-}$.

concentric arrangement of metal and chalcogenide, to the Cu_{70} species, which is disk-like and has the bulk, layered structure. But each material has a different crossover size: for silver chalcogenides, for instance, the switch occurs at the Ag_{114} cluster.

Large inorganic clusters are nothing new. Metal-oxide ions of the form $M_{12}O_{40}^{n-}$ were identified about 170 years ago by Berzelius, although their structures were not known until over a century later. These polyoxometalates, formed by early transition metals such as vanadium, molybdenum and tungsten, are hierarchical assemblies composed of MO_6 octahedral units linked into higher-order structures by sharing oxygen atoms at edges and vertices (Fig. 1). Their self-assembly is pH-dependent: increasingly acidic conditions drive growth towards the metal oxide, and basic conditions favour the discrete octahedral anions. Species as large as

$W_{148}O_{524}$ can be prepared in a clean synthesis from the corresponding metal salts at the appropriate pH (M. T. Pope, Georgetown Univ.).

These giant anions have been used as building blocks for higher-order assemblies by functionalizing their surfaces with organic groups. Octahedra are plucked off one by one to create reactive defects, and organic groups are attached to these sites via bridging metal atoms: an ester via tin; a porphyrin via niobium. The latter generates a chromophore coupled to a redox-active and strongly acidic group, with the promise of interesting photoelectrochemistry. The prospect of using traditional organic synthesis to assemble complex inorganic architectures is an indication that traditional subdivisions in chemistry are now largely redundant (E. C. Constable, Univ. Basel).

Most polyoxometalates form dense, ball-like clusters, but in contrast, the $\{Mo_{176}\}$ cluster is a giant wheel (A. Müller, Univ. Bielefeld; also *Angew. Chem. Int. Edn Engl.* **37**, 1220–1223; 1998). Now the wheel has been given hub caps in the form of $\{Mo_{36}\}$ clusters, which apparently form on the template of the wheel's inner rim — these smaller species are not stable on their own in solution. Such templating suggests a means of leap-frogging to large, complex assemblies via smaller intermediates.

Indeed, this principle has been extended to organize polyoxometalates into infinite ordered arrays, by using a crystal surface as the template. On the surface of silver, the silicotungstate ion $SiW_{12}O_{40}^{4-}$ forms a monolayer when deposited from solution (W. Klemperer, Univ. Illinois). The close-packed array of ions is visible under a scanning tunnelling microscope as a nearly square

(in fact, slightly rhombic) grid. Because the underlying metal lattice is hexagonal, the monolayer can't fit perfectly, but the ions can be arranged on the surface so that the four 'footprints' of the lowermost oxygen atoms maintain a largely consistent pattern of binding sites.

The same approach can be used to organize polyhedral borane ions (M. F. Hawthorne, UCLA) on a gold surface. A sulphhydryl group (SH) is substituted for a hydrogen atom in the cage-like $closo-B_{12}H_{12}$, rendering this borane cluster suitable for attachment to gold, and surface organization takes place (L. J. Yeager *et al.*, *J. Am. Chem. Soc.* **120**, 9961–9962; 1998). This kind of precise marriage between inorganic materials might be used to make thin-film interfaces with atom-scale definition. Other techniques for doing this, such as chemical vapour deposition, do not yet offer comparable control, as they rely on nucleation at random impurity sites.

Inorganic clusters can also be controlled with organic chemistry. Long-chain hydrophobic molecules with a hydrophilic, metal-binding head at each end (bola-amphiphiles) will prostrate themselves at an air–water interface in stacked layers (one-headed amphiphiles, in contrast, stand upright, tails first). The molecules then bind metal ions between adjacent ends, and this arrangement can be converted into a superlattice of equidistant inorganic clusters whose spacing is determined by the chain length (M. Lahav, Weizmann Inst.). Self-assembling lipid aggregates can also be used as templates to arrange metal clusters in complex patterns: helical tubules made of tapes of chiral lipid bilayers will support a helical array of gold particles deposited by reduction of a

Auctions

Darwin and Archimedes come under the hammer

On 11 November at Christie's, London, some Darwiniana comes up for sale — in the form of a signed photograph of the great man, taken at Down House in 1881, the year before his death; a sketch of primrose anatomy made by Darwin; and letters by him and other Victorian scientific worthies (T. H. Huxley *et al.*). The photograph comes with one of J. S. Henslow, Darwin's mentor when he was at the University of Cambridge; minimum guide price is £2,500 (US\$4,200). Autograph hunters can also bag Gregor Mendel's signature (£500), which appears on a bill from his bookseller.

But most items at this auction of natural history are botanical, zoological and ornithological illustrations. Examples of the work of the greatest of artists are available, among them F. L. Bauer, G. D. Ehret, P.-J. Redouté, John Gould and

Edward Lear (a copy of *Illustrations of the Family of Psittacidae*, of which only 175 were printed, has a guide price of £65,000).

Less familiar to Western eyes are the gems of paintings shown here, which come from an album of Chinese watercolours of 50 marine animals and monsters, produced in the mid-eighteenth century (£60,000).

Meantime, on 29 October at Christie's, New York, a canonical piece of intellectual history is on offer with the auction of the only remaining copy of the text on which accounts of some of Archimedes' ideas are based — including Archimedes' Principle itself. The parchment was produced in Constantinople in the tenth century, and has been overwritten with Greek religious texts. Nonetheless, the guide price is \$800,000–\$1,200,000.

Visit www.christies.com to find details of both auctions.

Tim Lincoln



gold salt (S. Mann, Univ. Bath).

Control of structure on still greater length scales, comparable to those in the natural organic–inorganic composites of bone, tooth enamel, spicules and nacre, might become possible if we can establish how nature does it (S. Mann; L. Addadi, Weizmann Inst.). In these structures, the crystallographic lattice is subjugated by a more powerful shaping agent, such as an organic membrane. This sort of process might be reproduced using organic, macromolecular

growth promoters or inhibitors at specific crystal faces. One way to achieve that is to use nature's own combinatorial resource, the immune system, to raise antibodies that recognize not organic antigens but the shape and form of particular crystal surfaces (L. Addadi; D. Perl-Treves *et al.*, *Chem. Biol.* **3**, 567–577; 1996).

Is this inorganic or organic chemistry, biology or materials science? Does it matter? □

Philip Ball is an associate editor of *Nature*.

Palaeontology

Forerunners of four legs

Philippe Janvier

Any textbook or popular book on palaeontology published since the 1930s tells us that the four-legged land vertebrates, or tetrapods, arose by the end of Devonian times (about 365 million years ago) from a group of lobe-finned fishes, the osteolepiforms. This is usually illustrated by a dramatic reconstruction of one of the best-known osteolepiforms, *Eusthenopteron foordi*, crawling out of a pond on its paired fins, to show how these could be the precursors of limbs. Given that this evolutionary picture is so widespread and popular, it must be based on well-corroborated palaeontological data. Thus, one may wonder why it appears as news in a paper by Per Ahlberg and Zerina Johanson¹ on page 792 of this issue.

The reason is that, until very recently, the theory of the osteolepiform–tetrapod relationship was thought to be well established on anatomical grounds, but the details remained extremely vague. For example, nobody could agree on whether osteolepiforms were merely an extinct fish group, related to modern-day tetrapods, or whether they included some species that were more closely related to tetrapods than others. Although it was generally accepted that the best-studied osteolepiform *E. foordi* shares some unique anatomical characters with the tetrapods, the status of the many other species referred to as osteolepiforms was not clear.

The theory of the osteolepiform–tetrapod relationships has come about by a complex route, which was documented by B. Schaeffer² and C. Patterson³. To cut a long story short, this theory arose in 1861 with the work of T. H. Huxley on his 'crossopterygian' fishes — an assortment of Palaeozoic fishes, including the living polypters (bichirs). Huxley considered crossopterygians as related to lungfishes which, in turn, were the favourite tetrapod ancestors for evolutionists of that time. In 1892, the American palaeontologist E. D. Cope extracted from Huxley's crossopterygians an ensemble of

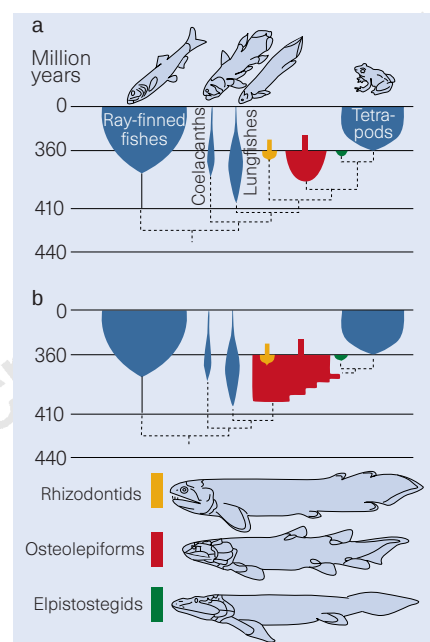


Figure 1 Inter-relationships of bony fishes and the four-legged vertebrates (tetrapods). **a**, The status of osteolepiforms (red) and rhizodontids (yellow) in most of the previous phylogenies. **b**, The revised status of osteolepiforms and rhizodontids according to the new analysis by Ahlberg and Johanson¹.

fishes which he called rhipidistians. These fishes were strikingly similar to the earliest tetrapods known at that time, the 330-million-year-old amphibians referred to as labyrinthodonts. The similarities included the pattern of the skull bones and the peculiar, folded structure of the teeth. Among Cope's rhipidistian fishes were osteolepiforms, such as *E. foordi*, from the 375-million-year-old Devonian rocks of Quebec.

Cope's ideas soon became widely accepted, and detailed studies on the anatomy of *E. foordi* yielded further evidence that it was ancestral to tetrapods. Like tetrapods, it had choanae — internal nostrils that allow it to breathe air when the mouth is closed — and a paired-fin skeleton that foreshadows the

tetrapod limb skeleton. Although many other Devonian and Carboniferous fossil fishes had a similar anatomy to this famous fossil, their overall aspect was somewhat different. In 1937, these were gathered with *E. foordi* into a group called the osteolepiforms (by reference to the earliest-described genus, *Osteolepis*, from the Devonian rocks of Scotland), on the basis of a few anatomical characters that seemed to be unique. Nevertheless, most discussions on tetrapod origins bore on comparisons with *Eusthenopteron* alone, and few palaeontologists have studied the other osteolepiforms in this respect.

In 1981, as the osteolepiform–tetrapod relationship was the received view, Rosen *et al.*⁴ published a provocative paper. They tried to show that the anatomical characters used for relating *E. foordi* (and all osteolepiforms) to tetrapods were merely primitive characters of bony fishes, and that lungfishes were the closest relatives of tetrapods — as was accepted during most of the nineteenth century. This triggered acrid responses from many palaeontologists, but can retrospectively be seen as a whiplash on the back of an old theory. After this controversy, a lively interest in tetrapod origins arose again, resulting in a number of discoveries that were often due to a reconsideration of old material. For instance, a group of Late Devonian fishes, the elpistostegids, were thought to be aberrant osteolepiforms. But these have been revisited, and it turns out that they are much closer to tetrapods than are the osteolepiforms. Thanks to the impressive number of Late Devonian tetrapods discovered during the past ten years, the gap between elpistostegids and tetrapods is now largely filled.

However, the tree down to the osteolepiforms still needed to be investigated. For want of a better solution, the osteolepiforms were regarded as being monophyletic — that is, a group of species that share some unique characters and all descend from a common ancestor that possessed at least one of these characters (Fig. 1a). Based on the most extensive character set ever used to analyse osteolepiform relationships, Ahlberg and Johanson¹ now provide evidence that osteolepiforms are not a monophyletic group. Their apparently unique characters are, in fact, transitory, and have been lost in the elpistostegids, tetrapods and rhizodontids (a fossil fish group that was long regarded as highly specialized osteolepiforms, and then as being more primitive than osteolepiforms). In other words, osteolepiforms are an ensemble of primitive tetrapodomorphs, a large group that also includes the rhizodontids, elpistostegids and tetrapods (Fig. 1b).

The most exciting consequence of this analysis rests in the base of Ahlberg and Johanson's tree. If osteolepiforms were

monophyletic, the discovery of increasingly primitive species of this group would merely have told us about the history of osteolepiforms. But because they now seem to be an ensemble of basal tetrapodomorphs, we can expect the discovery of early and primitive forms (such as *Kenichthys* from China) to tell us about the early history of the entire lineage that leads, ultimately, to the tetrapods. The aim is to arrive closer and closer to the point of divergence between this lineage (the

tetrapodomorphs) and that leading to the lungfishes (the dipnomorphs). □

Philippe Janvier is in the Laboratoire de Paléontologie, Muséum d'Histoire Naturelle, URA 12 CNRS, 8 Rue Buffon, 75005 Paris, France. e-mail: janvier@cimrs1.mnhn.fr

1. Ahlberg, P. E. & Johanson, Z. *Nature* **395**, 792–794 (1998).
2. Schaeffer, B. *Syst. Zool.* **14**, 115–118 (1965).
3. Patterson, C. in *Syst. Assoc. Spec. Vol. 15* (ed. Panchen, A. L.) 159–175 (Academic, London, 1980).
4. Rosen, D. E., Forey, P. L., Gardiner, B. G. & Patterson, C. *Bull. Am. Mus. Nat. Hist.* **167**, 159–276 (1981).

Planetary science

Oceans inside Jupiter's moons

Fritz Neubauer

For almost 50 years, exploration of the large Galilean moons of Jupiter has delivered a continuing stream of surprises. The four moons — Io, Europa, Ganymede and Callisto — have increasingly come to appear as an exotic and remarkably diverse planetary system in their own right. In 1979 we had the discovery of the most dramatic volcanism known in the Solar System, on the innermost Galilean satellite Io, during the Voyager space mission¹; and in 1996 the Galileo spacecraft revealed that Ganymede has a strong magnetic field², probably stemming from an internal dynamo.

Now, in the paper on page 777 of this issue³, comes strong evidence for the existence of oceans not very far below the surfaces of Europa and Callisto. This evidence results from the observation of electromagnetic induction signatures^{3,4} by the Galileo magnetometer experiment, which indicates that the interiors of these satellites are electrically conducting. The likeliest explanation is that the conductor is liquid water containing some electrolyte.

Over the years, ideas about the insides of Jupiter's satellites have evolved from being largely speculation to providing a mature physical picture. From the low bulk densities of Europa, Ganymede and Callisto, combined with arguments about the origin of the Solar System, it became clear early on that these satellites must contain substantial amounts of water ice or liquid water, as well as silicates and iron. By contrast, water is absent from Io, which is consistent with the violent volcanism there.

When the observations from the Voyager mission became available, thermal models⁵ of the satellite interiors, describing the internal structure of these bodies through time, were sufficiently constrained for the first time. The observations included data on the satellites' gravitational field, on their size and shape, and on surface geology. Important inputs into the thermal models are the internal heat sources of each satellite. It turned out from the modelling work that the interi-

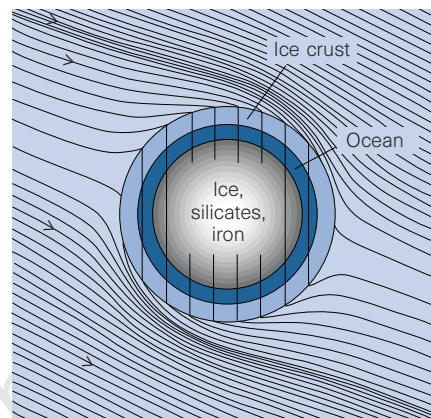


Figure 1 Interior of a satellite with a subsurface ocean. Because of the electrical conductivity of the ocean, conferred by its electrolyte content, the time-varying horizontal (equatorial) component of the locally homogeneous planetary magnetic field is pushed out, whereas the constant vertical component is unaffected. The depth of the ocean is exaggerated for clarity.

ors of Europa and Ganymede must be differentiated — that is, layered, with material becoming increasingly dense towards the centre as a consequence of gravitational separation after melting. It also became clear that, under certain conditions, liquid water could exist in the outer mantle⁶.

In these early models, Callisto was considered to be undifferentiated. The improved determination of the satellites' gravity fields by the Galileo mission, however, not only confirmed the differentiation of Europa⁷ and Ganymede⁸ but suggested some limited differentiation of Callisto's interior⁹. Independent evidence for the existence of a subsurface ocean at Europa came from surface observations by the Galileo imaging experiment^{10,11}, which provided indications that icebergs had drifted across Europa's surface at some time in the past.

So we had hints that subsurface oceans existed in the geological history of some satellites, but no conclusive evidence that oceans are present today. Magnetic field observations made during several close fly-

bys of Europa and Callisto by the Galileo spacecraft have, however, provided independent data showing that oceans are indeed present now on Europa and Callisto. The data were obtained by a completely different method, electromagnetic induction, from those used previously.

Electromagnetic induction techniques probe diverse targets and are important diagnostic tools for geophysicists studying the Earth. In principle, time-varying natural or artificial primary magnetic fields generate electric fields by induction according to Faraday's law. If the region of space to be explored contains materials with sufficient electrical conductivity, currents are generated which produce secondary magnetic fields. So these secondary magnetic fields can be used to investigate the presence of electrically conducting materials.

In the case of Jupiter, the primary field is the magnetic field from the interior of the planet modified by that generated by particles in the radiation belt³. Because this magnetic field is asymmetric with respect to Jupiter's rotational axis, the fast rotation of the planet (period of ten hours), combined with the slow orbital motion of the satellites, causes sizeable temporal variation in the primary field at the satellite. Now, for the extreme case of high electrical conductivity, the temporally varying portion of the primary field is excluded completely from the conducting interior.

In fact, the magnetic signatures observed at Callisto and Europa are quite strong, and require a conducting volume extending almost to the surface of the satellites. The interpretation is complicated by the simultaneous presence of magnetic fields generated in the surrounding plasma⁴, and by internal magnetic fields due to a dynamo or natural remanent magnetization. The latter can be distinguished from the induced field, however, because they are rigidly 'fixed' to the satellite, whereas the induced field follows the varying magnetic field of Jupiter.

The final step in this chain of reasoning is the realization that the high electrical conductivity must occur in the outer part of the satellite. Here the most straightforward — but not the only — explanation is that this high electrical conductivity is the electrolytic conductivity of a subsurface ocean. As in sea water on Earth, the electrolytes may come from leaching from 'rocky' material.

Whereas the existence of a subsurface ocean surmised in this way did not come as a complete surprise in the case of Europa, it stretches models of the interior and thermal evolution of Callisto. Ganymede, too, may also have a subsurface ocean, although observation of its induction signature is hampered by the satellite's strong dynamo field. On a more speculative note, it has been suggested that life may exist or have existed in the subsurface oceans. A Europa orbiter

mission is being considered by NASA to study them further¹². □

Fritz Neubauer is at the Institut für Geophysik und Meteorologie, Universität Köln, 50923 Köln, Germany.

e-mail: neubauer@geo.uni-koeln.de

1. Morabito, L. A., Synnott, S. P., Kupferman, P. N. & Collins, S. A. *Science* **204**, 972 (1979).
2. Kivelson, M. G. et al. *Nature* **384**, 537–541 (1996).
3. Khurana, K. K. et al. *Nature* **395**, 777–780 (1998).
4. Neubauer, F. M. J. *Geophys. Res.* **103**, 19843–19866 (1998).

5. Schubert, G., Spohn, T. & Reynolds, R. in *Satellites* (eds Burns, J. A. & Matthews, M. S.) 224–292 (Univ. Arizona Press, Tucson, 1986).
6. Cassen, P. M., Reynolds, R. T. & Peale, S. J. *Geophys. Res. Lett.* **6**, 731–734 (1979).
7. Anderson, J. D. et al. *Science* **281**, 2019–2022 (1998).
8. Anderson, J. D., Lau, E. L., Sjogren, W. L., Schubert, G. & Moore, W. B. *Nature* **384**, 541–543 (1996).
9. Anderson, J. D. et al. *Science* **280**, 1573–1576 (1998).
10. Carr, M. H. et al. *Nature* **391**, 363–365 (1998).
11. Pappalardo, R. T. et al. *Nature* **391**, 365–368 (1998).
12. http://www.jpl.nasa.gov/ice_fire/EO_Info.htm

Evolutionary biology

The geometry of adaptation

Nick Barton

Darwin¹ explained the evolution of complex adaptations, such as the vertebrate eye, through the gradual “accumulation of innumerable slight variations, each good for the individual possessor”. In contrast, the early geneticists emphasized the role of major mutations, and disregarded natural selection². The mathematical theory of Fisher, Haldane and Wright reconciled Mendelian genetics with Darwinian gradualism by showing that genetic variants that slightly enhanced reproductive success would quickly replace their competitors. Nevertheless, the conflict over the magnitude of the genetic changes responsible for adaptation continues to this day³. Although classical population genetics can find the consequences of a given pattern of selection, it says nothing about the causes of such selection. Indeed, theoretical support for gradualism rests mainly on a geometrical analogy, proposed by Fisher⁴ in 1930, which suggests that large changes to a complex system are unlikely to be favourable. Now, reporting in *Evolution*, H. Allen Orr⁵ shows that adaptation may be based on larger factors than previously supposed.

Fisher's argument is simple. Suppose that an organism is described by n traits, and that its fitness increases towards some optimal value. If a population lies away from the optimum, then a random change that takes it inside an n -dimensional sphere centred on the optimum will increase its fitness, and may be established by natural selection (Fig. 1, overleaf). A small change is as likely to be favourable as unfavourable, whereas a change larger than the diameter, d , of the sphere must be disadvantageous. Fisher showed that, for large n , the chance that a change of magnitude r (a sum over all traits) increases fitness is $(1 - \Phi(x))$, where Φ is the cumulative normal distribution, and $x = r\sqrt{n}/d$ is a measure of the effective size of the change. So, as the complexity of the organism (represented by n) increases, a change of absolute magnitude r is less likely to be an improvement.

But a fundamental flaw in Fisher's argument was found, half a century later. Kimura⁶ pointed out that although small changes are

most likely to increase fitness, they are also most likely to be lost by chance. Even in a large population, a single mutation that increases relative fitness by s has a chance of only around $2s$ of being established⁴. Kimura showed that if the selective advantage of a mutation is proportional to its effective size, x , then the chance that the mutant is favourable and becomes common is $2x(1 - \Phi(x))$. The expected effective size of successful mutations is $E[x] = 4(\sqrt{2/\pi})/3 = 1.06$. Thus, mutations of intermediate effect are most likely to contribute. For example, with $n = 50$ dimensions, the first adaptive step will average 30% of the distance to the optimum, $d/2$. However, because its direction is random, it will take the population only about 4% further towards the optimum.

Orr⁵ now goes beyond these arguments by considering the whole evolutionary sequence by which a population approaches the optimum. The distribution of sizes of the first step is given by Kimura's formula. The population is then, on average, a factor $(1 - 2E[x])/n$ closer to the optimum. The process is then repeated from this new starting point. Orr shows that, to a close approximation, substitutions continue with a distribution that retains the same form, but has a continually shrinking scale (Fig. 1). Remarkably, the overall distribution of sizes of factors fixed over this adaptive ‘walk’ to the optimum is exponential, $\sim \exp(-\lambda x)$, with $\lambda = 2.9$. This result is independent of the dimension n , but does assume a uniform distribution of mutational effects.

Orr shows that for a variety of mutational distributions, biased either to smaller or larger effects, the distribution of factors fixed remains close to exponential (although with larger or smaller λ). Evolution by a series of gradually diminishing random draws means that the *largest* step, $x_{\max}$, is probably larger than the first, and increases logarithmically with n ; for example, with $n = 50$ dimensions, $E[x_{\max}] = 1.68$, corresponding to a magnitude 48% of the original distance to the optimum. Thus, large changes are established in the early stages of adaptation towards a fixed optimum, mainly because alleles with the



100 YEARS AGO

What was the condition of England in 1845? Her universities had degenerated into *hauts lycées*. With regard to the University teaching, I may state that even as late as the late fifties a senior wrangler – I had the story from himself – came to London from Cambridge expressly to walk about the streets to study crystals, prisms and the like in the optician's windows. Of laboratories in the universities there were none; of science teaching in the schools there was none; there was no organisation for training science teachers. If an artisan wished to improve his knowledge he had only the moribund Mechanics' Institutes to fall back upon. ... We lacked then everything which Germany had equipped herself with in the matter of scientific industries. Did this matter? Was it more than a mere abstract question of a want of perfection? It mattered very much! ... At this time we had, fortunately for us, in England, in very high place, a German fully educated by all that could be learned at one of the best equipped modern German Universities, where he studied both science and the fine arts. I refer to the Prince Consort. From that year to his death he was the foundation of our English educational renaissance ... From *Nature* 20 October 1898.

50 YEARS AGO

Co-operation between the Standard Oil Co. and its subsidiaries of America and the I.G. Farben-industrie of Germany led to the formation in 1929 of a joint company ... The final result of this co-operation brought to America the technique of the preparation of Buna S, an all-purpose synthetic rubber ... Howard tells with what sense of urgency he and his colleagues regarded the necessity for the setting up of a synthetic rubber industry, and how, in the face of a general apathy, his Company decided to erect the necessary plant for the manufacture of Buna S on a small scale. Even after Pearl Harbour, all their efforts to get this vital material made on a large scale came to nothing ... until the setting up of the Baruch Committee in August 1942; and the resulting appointment of a rubber director with absolute power cleared away all the difficulties, with the result that in a very short time the American synthetic rubber industry was among the largest in the country. From *Nature* 23 October 1948.

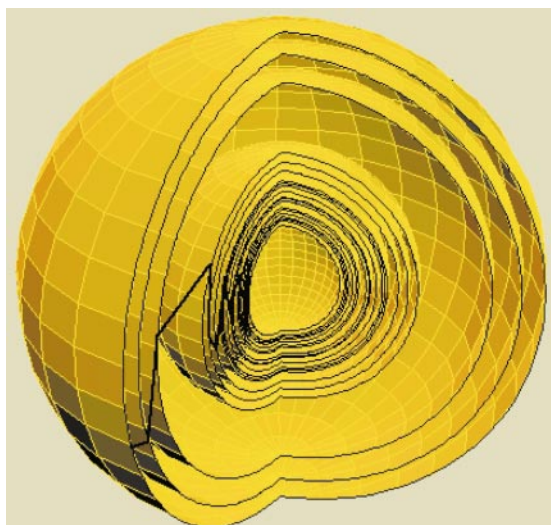


Figure 1 Approach of a population towards an optimum through selection of a series of random mutations, through a space of ten dimensions. The population begins at a point on the outer sphere, at a distance $d/2$ from the central optimum (where d is the diameter of the sphere). The first successful mutation has magnitude $r = 0.137d$, and takes the population 8.7% of the way to the optimum (first line, leading to the second sphere). The third successful mutation has the largest magnitude, $0.271d$; it is followed by smaller steps which, on average, follow a geometric series.

largest effect on fitness are most likely to be picked up by natural selection. Later, small changes do contribute, but they are not responsible for much of the improvement in fitness.

The prediction that the effects of the genes responsible for adaptation should follow an exponential distribution roughly applies to individual traits, as well as to the sum over all traits. Orr argues that studies⁷ of the quantitative trait loci (QTL), which distinguish close taxa, are consistent with this prediction. But it is not clear what would be predicted by other models of evolution, or whether these could be distinguished. Existing methods for detecting QTL are biased towards finding genes of spuriously large effect⁷. More convincing evidence for the importance of major genes comes from studies of 'candidate loci', which often contribute to divergence and standing variation^{8,9}. However, such examples hardly fit with Fisher's model, whereby mutations are drawn randomly from any of the many genes that might influence a trait.

Some functions may necessarily be composed of many minor changes. For example, the efficiency and accuracy of translation can be improved by using triplet codons that match the most abundant transfer RNAs. Each such change could have only a tiny effect on fitness¹⁰, so, for codon-usage bias, adaptation is based on minute variations. Nevertheless, such examples are hard to find and there are many limits to the success of weakly selected variants, beyond the low fixation probability identified by Kimura⁶ and Orr⁵. In a large, asexual population, a favourable allele can fix only if it confers an advantage that outweighs the loss in fitness owing to the deleterious alleles with which it arises⁴. Even with sexual reproduction, linkage reduces the effectiveness of natural selection, and can greatly reduce the fixation probability of a weakly favoured allele¹¹. Indeed, bias in codon usage is weaker in regions of reduced recombination in *Drosophila*¹².

Fisher's geometrical analogy is an extreme abstraction of organismic complexity. One problem is that it is hard to know what is meant by the number of dimensions, n . The morphology, development and behaviour of an organism can be described by an indefinite number of phenotypic dimensions that may all be influenced by many genes. The number of degrees of freedom, n , is presumably limited by the number of genetic loci ($\sim 10^5$ genes and $\sim 3 \times 10^9$ base pairs in man). However, once one thinks in terms of gene sequences rather than continuous traits, a different geometry is required. Although many sequences are possible (4^k with k bases), only a few ($\sim 3k$) can be reached by a single mutation. It may be possible to do an analysis similar to Orr's for this tightly restricted geometry¹³, although this might come to different conclusions. The issues raised by Orr's work are relevant to the optimization of any complex structure — be it computer software or a mechanical device. But the universal genetic code, together with simple mathematics of natural selection, may allow population genetics to provide a more general understanding of adaptation than is possible in other contexts. □

Nick Barton is in the Institute of Cell, Animal and Population Biology, King's Buildings, University of Edinburgh, Edinburgh EH9 3JT, UK.
e-mail: n.barton@ed.ac.uk

1. Darwin, C. *On the Origin of Species by Means of Natural Selection* (Murray, London, 1859).
2. Provine, W. *The Origins of Theoretical Population Genetics* (Univ. Chicago Press, 1971).
3. Orr, H. A. & Coyne, J. A. *Am. Nat.* **140**, 725–742 (1992).
4. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, 1930).
5. Orr, H. A. *Evolution* **52**, 935–949 (1998).
6. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, 1983).
7. Lynch, M. & Walsh, J. B. *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, MA, 1998).
8. McKenzie, J. A. & Batterham, P. *Trends Ecol. Evol.* **9**, 166–170 (1994).
9. Long, A. D., Mullaney, S. L., Mackay, T. F. C. & Langley, C. H. *Genetics* **144**, 1497–1510 (1997).
10. Bulmer, M. *Genetics* **129**, 897–907 (1991).
11. Barton, N. H. *Genetics* **140**, 821–841 (1995).
12. Kliman, R. M. & Hey, J. *Genetics* **137**, 1049–1056 (1994).
13. Gillespie, J. H. *Evolution* **38**, 1116–1129 (1984).

Daedalus

Dead metallic silence

Noise, that curse of modern living, is spread in many ways. The racket of a diesel engine, for example, is greatly amplified by vibrating sheet-metal surfaces such as the rocker cover, the sump-case and the body panels of the vehicle housing it. Workplace machines, domestic gadgets, and even clattering pots and pans, also owe much of their characteristic noise to the undamped vibration of metal surfaces. If all our products could be made (say) of wood, life would be much more peaceful.

So Daedalus wants to load engineering metals with vibration-damping inclusions. Cast iron, for example, one of the quietest of metals, contains many tiny graphite inclusions which absorb vibration. Sadly, they also act as crack-initiating sites, and make it brittle. Daedalus recalls that polystyrene (which has rather a metallic ring itself) is damped by certain organo-metallic ring compounds. The organic ring in these molecules can rotate stepwise around the bond joining it to the metal. The polymer is strongly damped in the acoustic range of frequencies corresponding to the stepping rate of the ring.

So DREADCO chemists are synthesizing organic ring molecules capable of binding to iron and other engineering metals, and are using powder sintering to disperse the most likely candidates into bulk metals. Cunningly, they are seeking molecules that are the stablest combination of that set of atoms. Trapped in its metallic cage at the sintering temperature, such a molecule might well decompose; but it would re-form again when the metal cooled. The resulting metals will be acoustically dead yet still tough and ductile — single molecules are far too small to act as crack initiators.

When perfected, DREADCO's 'Dead Metal' range of alloys should sweep through society. The big noises, such as diesel engines, are the obvious targets. But Daedalus attributes much of the edginess of modern living to the quieter but ever-present background noise from central-heating systems, air-conditioning units, office equipment, refrigerators, cookers, washing machines and so on. The formless, ever-changing insistence of this endless accompaniment to life gets on our nerves in some very subtle way. He recalls the story of a housewife in a modern kitchen who, hearing a sudden change in its complex background racket, asked in bewilderment, "What's just stopped?" With Dead Metal, they will all stop.

David Jones

Akeley's Africa

The size and elaborate staging of the great dioramas, such as Carl Akeley's gorillas, can still impress. Their cost upset museum authorities, but they aided conservation by promoting the idea of wildlife in its natural habitat.

Martin Kemp

Even in this age of multimedia extravaganzas, the most brilliant of the old habitat dioramas in museums of natural history still attract enthralled viewers. Tableaux of stuffed animals in vivacious poses are re-naturalized in panoramas dense with biological and geological detail, the 'real' foregrounds merging indiscernibly into deep backgrounds painted by specialist artists in shallow niches.

The concept was perfected by the Swedish taxidermist and sportsman, Gustaf Kolthoff. His Biological Museum, which opened in Stockholm in 1893, stands as the supreme surviving example of a habitat panorama. From two viewing platforms the spectator is presented with almost 4,000 zoological specimens, ecologically integrated within eight successive landscape types, the backgrounds of which were painted by Bruno Liljefors, Sweden's supreme nature artist. Scientific demonstration and popular appeal coexist in perfect harmony.

Outside Sweden, the ecological diorama was adopted enthusiastically in America. The diorama halls in the American Museum of Natural History in New York, into which huge resources were poured in the first half of this century, vividly convey the American vision of the sublime wilderness, the definitive expression of which was the African Hall of Carl Akeley.

A taxidermist who devised new techniques for naturalistic animation, Akeley gained public renown as an intrepid explorer, environmental conservationist, ani-



Carl Akeley and Albert Butler's "Gorilla group", with landscape by William Leigh, in the African Hall at the American Museum of Natural History, New York. Photo by Finnin/Beckett.

mal sculptor and inventor, who was said to have killed a leopard in hand-to-paw conflict. It was on his return in 1911 from an expedition to hunt animals for a dynamic tableau of alarmed elephants that Akeley conceived the African Hall, within which "The Alarm" still stands. The grand hall of luminous dioramas, in a framework of rich wood decorated with sculptural reliefs, became the great vehicle for his "biography of untouched Africa", conveying his mission to conserve the wonderful wildlife threatened by human encroachment.

Mountain gorillas, the subject of a dramatic group in a sublime landscape, were in particularly urgent need of protection from

hunting. Akeley's 1922 proposal to the Belgian government to save the gorillas led to the establishment of the first national park in the Congo. The foreground of the diorama contains more than 75,000 laboriously mounted artificial flowers and leaves, while the scenery — hailed by Akeley as beyond compare — was painted by William Leigh, a specialist in panoramas.

A photograph by Akeley portrays the expeditionary force, including Leigh and a fellow painter, armed with brushes, paints, canvases, easels and a small-scale diorama box. Meticulous care was taken in assembling field data on every aspect of the flora, fauna, geology and meteorology of the setting. The goal was, in Leigh's words, "complete pictures, faultless history, perfect science".

Yet the costs incurred by the expeditions and assembly of the dioramas left the museum's education department open to criticism from scientific curators who believed that the museum's primary role should be research — not public displays "totally devoid of intellectual content". Such claims failed to recognize the role the dioramas had played in persuading both biologists and the public of the need to view every living thing as part of the rich and evolving panorama of nature's complexity. □

Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.

Quotes from Karen Wonders, *Habitat Dioramas*, Uppsala University Press, Uppsala (1993).



Akeley's photograph of William Leigh (standing) and A. A. Jansson, making studies for the African lion diorama, 1926–27.

Neuron loss in APP transgenic mice

Alzheimer's disease (AD) is a progressive neurodegenerative disorder that affects a large proportion of the elderly population. Amyloid plaques, which are a neuropathological characteristic of AD, have been reproduced in transgenic mice by the overexpression of mutant forms of the amyloid- β precursor protein (APP) known to cause familial AD. Here we report that these mice exhibit selective neuronal death in the brain regions that are most affected in AD, suggesting that amyloid plaque formation is directly involved in AD neuron loss.

APP23 mice express mutant human APP (the 'Swedish mutation'), and develop amyloid- β peptide (A β)-immunoreactive plaques, primarily in neocortex and hippocampus, progressively with age¹. The plaques have most of the characteristics of human AD plaques, including fibrillar A β cores, and are surrounded by dystrophic neurites, activated microglia and astrocytes¹.

Here we have used APP23 transgenic mice 14–18 months old and confirmed that there was robust and prominent plaque deposition throughout the entire neocortex and hippocampus (Fig. 1). Staining of adjacent sections for A β and Congo red revealed that more than 90% of the plaques had a dense, fibrillar amyloid core, typical of human amyloid plaques (Fig. 1). Diffuse plaques also occurred, but were generally

limited to specific subregions and were most frequently found in animals with a high plaque burden. Vascular amyloid was also present and occurred most frequently in meningeal, neocortical and thalamic vessels.

Quantification of amyloid plaque load in neocortex and hippocampal area CA1 of our transgenic mice showed that A β -immunoreactive plaques occupy 1–26% and 0.1–28% of the total volume of neocortex and CA1, respectively (see Fig. 2b,d). Plaque loads in neocortex and CA1 were highly correlated ($r=0.95$, $P<0.0001$). The wide range of plaque load reflects our use of both hemi- and homozygous mice, and facilitated correlative analysis between plaques and neurons. By using unbiased stereological methods, we estimated the total number of neurons in both the CA1 pyramidal layer and neocortex. The number of CA1 pyramidal neurons was significantly decreased in the transgenic mice compared with controls (Fig. 2a), and was inversely correlated with CA1 plaque load in the transgenic mice (Fig. 2b). Overall, in the transgenic mice, CA1 neuron loss was 14% and reached 25% in mice with a high plaque load. A thinning of the pyramidal layer was clearly evident (Fig. 2e,f,i,j), and the neurons appeared to be interrupted in the vicinity of plaques. Pyramidal neurons

around amyloid plaques were occasionally pyknotic and/or exhibited clumps of condensed chromatin and irregular membrane structure (Fig. 2k), morphology typical of dying neurons.

No quantitative evidence of neuron loss was observed in the neocortex as a whole, even in the most severely affected animals. Mean values of the total number of neurons were similar in the control and transgenic groups (Fig. 2c), and no relationship was seen between plaque load and neuron number in the transgenic mice (Fig. 2d). In most regions, neurons around cortical plaques appeared to be physically displaced as plaques formed (Fig. 2e,f). However, in subregions such as the piriform and entorhinal cortices, qualitative evidence of neuron loss was clearly observed (Fig. 2g,h). Throughout the neocortex, plaques caused severe disruption of the normal cytoarchitecture.

The cause, extent and localization of neuron loss in AD have been controversial. However, assumption-free stereological techniques have been applied in human brain, and the results coincide with our findings in transgenic mice. Selective neuron loss in area CA1 of the hippocampus has been demonstrated in AD². In the neocortex, no global loss of neurons has been observed³, although degeneration in selected neocortical regions undoubtedly occurs⁴. These deficits occur in regions that are not severely affected by normal ageing in either humans^{2,5} or B6 mice⁶. The mechanism for the selective vulnerability of CA1 neurons and subpopulations of cortical neurons in both AD and our mouse model remains unclear.

Structural studies in two other APP-transgenic mice known to develop amyloid plaques — the PD-APP mouse⁷ (Stock-TgN(PDGF-APP_{V717F})) and the PrP-APP mouse⁸ (B6.SJL-TgN(PrP-APP_{K670N/M671L})) — have reported no significant neuron loss in hippocampal and cortical subregions^{9,10}. For the PD-APP mouse, this difference may be because many of the plaques are of the diffuse type with little impact on neuronal cytoarchitecture, because of differences in the APP isoform/mutation expressed in a particular mouse strain, or because only part of the CA1 field was sampled⁹. In PrP-APP mice, a slight non-significant decrease in CA1 neuron number was reported compared with controls¹⁰. This result is consistent with our data in that the subset of APP23 mice with a similar CA1 plaque load (0–5%; see Fig. 2b) also exhibited a non-significant loss of CA1 neurons.

We have shown that the formation of amyloid plaques can lead to region-specific

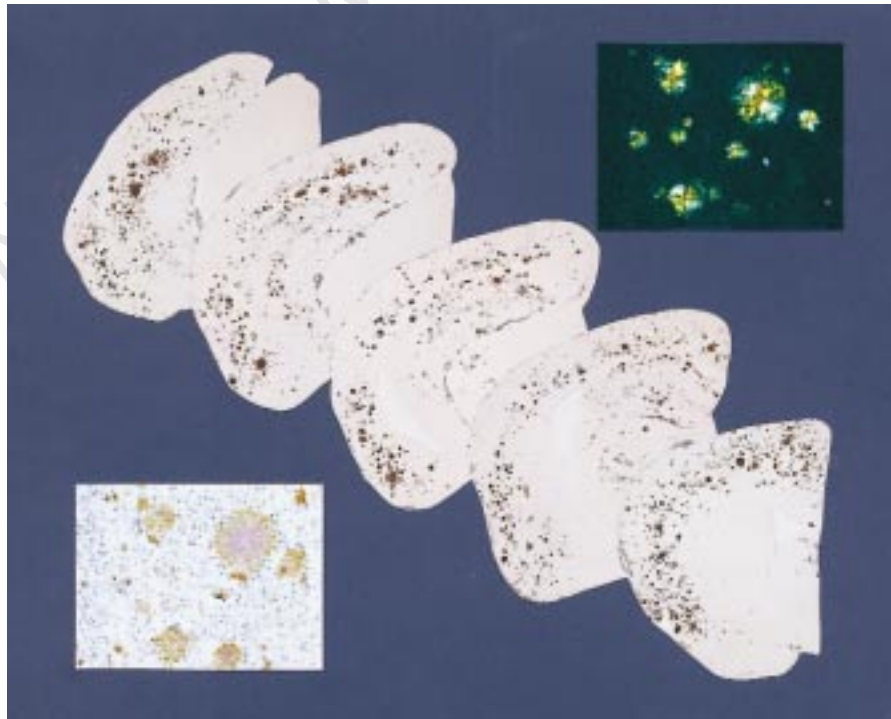


Figure 1 Distribution of amyloid plaques in a 14-month-old APP23 transgenic mouse (B6.D2-TgN(Thy1-APP_{K670N/M671L})). Immunostaining with the NT-11 A β antibody¹ shows amyloid plaques most frequently in neocortex and hippocampus. Insets, adjacent sections stained for A β (bottom left) and Congo red (top right). The plaque load in this mouse was 15% for the neocortex and 12% for area CA1 of the hippocampus, approximately the mean of all transgenic animals analysed.

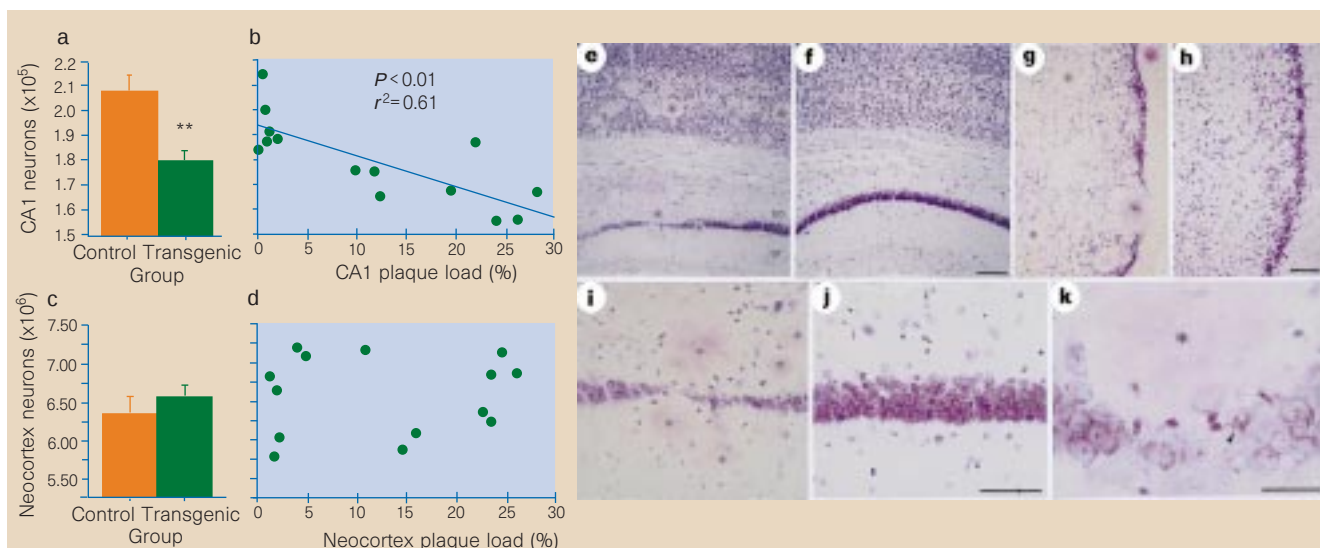


Figure 2 Region-specific neuron loss in APP23 transgenic mice. **a**, A significant ($P < 0.01$) decrease in the number of CA1 pyramidal neurons was seen in APP transgenic mice compared with control mice. **b**, The loss of CA1 neurons correlated with plaque load. **c**, **d**, In neocortex, no global loss of neocortical neurons (**c**) or relationship to plaque load (**d**) was observed. **e–k**, Cresyl violet staining reveals disrupted neuronal cytoarchitecture and a region-specific loss of neurons. The deeper layers of the parietal/somatosensory cortex and CA1 of the dorsal hippocampus of an APP-transgenic (**e**) and a control (**f**) mouse are shown. The pyramidal layer (**p**) appears thinner and interrupted near plaques (asterisks); in the parietal cortex, neurons near plaques often appear physically displaced. In selected cortical subregions, such as the entorhinal/piriform cortex, neurons around amyloid plaques appear to be lost in APP transgenic (**g**) compared with control (**h**) mice. A thinning of the CA1 pyramidal cell layer was observed in APP transgenic (**i**) compared with a control (**j**) mouse. **k**, High-power

view of a plaque interrupting the CA1 pyramidal layer with at least one cell exhibiting morphology typical of a dying neuron (arrowhead). Abbreviations: so, stratum oriens; sr, stratum radiatum. Scale bars: 100 μm (**e–h**); 50 μm (**i, j**) and 20 μm (**k**). Mice (14 hemi- and homozygous transgenic, and 10 littermate controls) were transcardially perfused with 4% paraformaldehyde and brains removed. The left cerebral hemisphere was processed for paraffin-embedding by standard methods. Serial coronal sections (microtome setting, 25 μm) were cut through the entire brain. Every tenth or twentieth section was selected with a random start for hippocampus and cortex, respectively, yielding 12–14 sections per region for each animal. Sections were stained with cresyl violet to visualize neuronal nuclei. Total neuron number was estimated¹⁴ by calculating the volume of neocortex and CA1 pyramidal cell layer according to the Cavalieri point-counting principle, and multiplying by neuronal density estimates in the same regions by counting neurons that came into focus within 100

optical disectors systematically spaced through each region. Disectors were 12 μm high (in an average actual section thickness of 29 μm , measured with an optical encoder attached to the microscope), and the frame area was 190 and 750 μm^2 for CA1 and cortex, respectively. Plaque load was estimated on an adjacent set of sections by sampling through the regions at $\times 20$ magnification and counting the percentage of points on an overlaid grid which were over plaques. For CA1, plaque load was measured in all layers (stratum oriens, pyramidal cell layer, stratum radiatum and stratum lacunosum-moleculare). Sampling was optimized to produce a coefficient of error (mean 7.6%) lower than the observed biological variability. All quantification was done with Stereologer software (SPA, Alexandria, VA). Reported numbers are for left hemisphere only. Student's *t*-test was used for group comparisons and linear regression with analysis of variance was used for analysing the relationship between plaque load and neuron number.

loss of neurons in a transgenic mouse. Neuron loss was observed primarily in the vicinity of plaques, but intraneuronal amyloidogenic APP processing cannot be excluded as an additional cause. The extent of the observed loss is less than that reported in end-stage AD², possibly because overexpression of APP has a neuroprotective effect¹¹. It is also likely that neuron loss would increase in APP23 mice with further ageing^{12,13}. Finally, species-specific factors may also be involved in the neurodegeneration seen in AD¹³. Taken together, these results demonstrate that the formation of amyloid plaques plays a primary role in AD neuron loss, and indicate that this mouse model will be useful for the evaluation of neurodegenerative mechanisms and therapeutic intervention.

Michael E. Calhoun*,
Karl-Heinz Wiederhold†,
Dorothee Abramowski†, **Amie L. Phinney***,
Alphonse Probst*,
Christine Sturchler-Pierrat†,

Matthias Staufenbiel†, **Bernd Sommer†**,
Mathias Jucker*.

*Neuropathology, Institute for Pathology,
University of Basel,
CH-4003 Basel, Switzerland
e-mail: mjucker@uhbs.ch

†Novartis Pharma Ltd, Nervous System Research,
CH-4002 Basel, Switzerland

1. Sturchler-Pierrat, C. et al. *Proc. Natl Acad. Sci. USA* **94**, 13287–13292 (1997).
2. West, M. J., Coleman, P. D., Flood, D. G. & Troncoso, J. C. *Lancet* **344**, 769–772 (1994).
3. Regeur, L., Jensen, G. B., Pakkenberg, H., Evans, S. M. & Pakkenberg, B. *Neurobiol. Aging* **15**, 347–352 (1994).
4. Gomez-Isla, T. et al. *J. Neurosci.* **16**, 4491–4500 (1996).
5. Pakkenberg, B. & Gundersen, H. J. *J. Comp. Neurol.* **384**, 312–320 (1997).
6. Calhoun, M. E. et al. *Neurobiol. Aging* (in the press).
7. Games, D. et al. *Nature* **373**, 523–527 (1995).
8. Hsiao, K. et al. *Science* **274**, 99–102 (1996).
9. Irizarry, M. C. et al. *J. Neurosci.* **17**, 7053–7059 (1997).
10. Irizarry, M. C., McNamara, M., Fedorchak, K., Hsiao, K. & Hyman, B. T. *J. Neuropathol. Exp. Neurol.* **56**, 965–973 (1997).
11. Mattson, M. P. et al. *Neuron* **10**, 243–254 (1993).
12. Jucker, M. & Ingram, D. K. *Behav. Brain Res.* **85**, 1–26 (1997).
13. Geula, C. et al. *Nature Med.* **4**, 827–831 (1998).
14. West, M. J. *Neurobiol. Aging* **14**, 275–285 (1993).

Why some papers have long citation lifetimes

When scientists publish their results they wonder whether their papers will be remembered for many years, as some are, or quickly forgotten. Here we show why the average lifetimes of papers in various scientific disciplines differ so much. We find that both the length of the paper and the speed of growth of the field are crucial in determining the duration and number of citations it receives.

All 165 papers published in the *Astrophysical Journal* (Astrophys. J.) in 1954 produced citations during the following 40 years that showed an average exponential decay with a half-life of 29.3 ± 1.7 years¹. Even after 40 years the citations per year were 39% of the initial numbers. Such a long lifetime seems surprising. Do other

Table 1 Half-lives of papers in different scientific disciplines

Journal (discipline)	Papers	Half-life (years)	Half-life corrected for growth (years)
<i>Astrophys. J.</i> (astronomy)	165	29.3	8.0
<i>J. Am. Chem. Soc.</i> (chemistry)	107	15.1	8.8
<i>J. Geophys. Res.</i> (geophysics)	111	6.4* 28.9†	5.5* 9.8†
<i>Phys. Rev.</i> (physics)	123	10.6	5.9
<i>Science</i> (general)	108	8.9	7.9
		Mean	7.6 ± 1.7

*Before 1970.

†After 1970.

sciences also have such a long lifetime?

A similar study was done on all 123 papers published in the January 1959 issues of the physics journal *Physical Review* (*Phys. Rev.*) and followed for 30 years in the Science Citation Index. Their citations showed a half-life of just 10.6 ± 0.4 years. Why is the average lifetime of papers in a leading physics journal so much shorter than for a leading astronomy journal?

Several differences between astronomy and physics come to mind. Astronomy is primarily an observational science, whereas physics is an experimental science. Astronomical objects are mostly permanent relative to human lifetimes, so papers invite comparisons with earlier results. But the dominant factor seems to be that astronomy has been growing more rapidly than physics in recent decades.

In the 40 years since 1954, *Astrophys. J.* grew from publishing 165 to 1,812 papers a year, a factor of 11.0. If there are 11 times as many papers being published now, it seems likely that an older paper has 11 times as much chance of being cited in a static field. Therefore, if the 1992 citations are reduced by that factor, the resulting half-life corrected to a static field becomes 8.0 years. Allowing for the growth in numbers of *Phys. Rev.* papers produces a similar corrected half-life of 5.9 years.

Papers published in 1959 in other sciences were considered, namely those in chemistry in the *Journal of the American Chemical Society* (*J. Am. Chem. Soc.*), geophysics in the *Journal of Geophysical Research* (*J. Geophys. Res.*) and general science in the Reports section of *Science*. Comparisons of these are fair because another study² has shown that the average numbers of references per paper in domestic and foreign journals in various sciences are a function only of the paper lengths.

Initially all of these had averages of 1 to 3 citations per paper per year. The results given in Table 1 show measured half-lives from 6.4 to 29.3 years, but after correction for journal growth they fall to within 5.5 and 9.8 years. If the numbers of members of the appropriate national societies are used as measures of growth, rather than the numbers of papers, the resulting corrected half-lives are similar but about 3.0 years longer. The wide range of lifetimes can

therefore be attributed primarily to differing growth rates in various sciences. Geophysics before and after 1970 shows the full range of this effect, resulting from an abrupt expansion in the late 1960s in the number of papers published.

Subfields of various sciences also show different lifetimes. For example, in *Astrophys. J.* the fraction of papers about cosmology and galaxies has increased from 6% in 1954 to 42% in 1992, a factor of 7. The growth rate for such papers is therefore not just a factor of 11.0, but of $7 \times 11.0 = 77$. This extra growth explains why the 1954 papers on galaxies and cosmology show a constant citation rate over 40 years. Papers in stellar astronomy, which have decreased relative to the total, show a shorter lifetime than the average of 29 years. Observational papers have a longer half-life (35 ± 2 years) than theoretical papers (22 ± 2 years).

Long papers produce more citations than short ones. In physics the average number of citations in 30 years is $7.4 + 6.7 \times$ (the number of pages), so for papers longer than about five pages, for which the first term is negligible, the number of citations is proportional to the length of the paper.

Of course, the importance and usefulness of a paper — two of the factors measured by citation counts — vary greatly from paper to paper. But on average, papers that are long and in sciences or subfields that are growing the fastest will be cited most often over the longest period.

Helmut A. Abt

Kitt Peak National Observatory,
Box 26732, Tucson, Arizona 85726-6732, USA

1. Abt, H. A. *Publ. Astron. Soc. Pacif.* **108**, 1059–1061 (1996).

2. Abt, H. A. *Publ. Astron. Soc. Pacif.* **99**, 1329–1332 (1987).

Nanodomain control in copolymer thin films

Diblock copolymers are composed of two chemically distinct chains covalently bonded together at one end. We have discovered how to control the orientation of diblock copolymer microdomains (made up of nanoscopic cylinders, or lamellae) in thin films. To do this, we eliminated all preferential interfacial segregation of the

components by anchoring random copolymers, composed of identical monomers to the diblock copolymer, to both interfaces of the film. The composition of the random copolymer is tailored to balance interactions between the segments of the copolymer and the interfaces.

Generally, the orientation of the microdomains of diblock copolymers in thin films is parallel to the surface because of the preferential segregation of one of the components to the interface¹. Removing these preferential interactions causes the copolymer microdomains to orientate perpendicular to the surface and substrate interfaces^{2–5}. Perpendicular alignment of the microdomains is favoured over the parallel arrangement because of configurational restrictions that force the diblock copolymer molecules to orientate parallel to non-preferential interfaces³.

As a model system to demonstrate this effect we used symmetric diblock copolymers of polystyrene and poly(methyl methacrylate), denoted as P(S-b-MMA), which have a lamellar morphology with a period of 36 nanometres. Random copolymers, denoted as P(S-r-MMA), which are composed of styrene and methyl methacrylate monomers randomly placed along the polymer chain and have a styrene fraction of 0.60, were used to produce the non-preferential surfaces. At this random copolymer concentration, the interfacial energies between each of the blocks with the random copolymer are equal⁶.

To anchor P(S-r-MMA) to the interfaces of the thin films, it was chemically grafted to a silicon substrate and affixed to the polymer–air interface by end-linking it to perfluorinated groups. The very low surface energy of the perfluorinated group forces it to the free surface, thereby anchoring the P(S-r-MMA) to the polymer–air interface. This is a simple way to confine a diblock copolymer, or mixtures of dissimilar polymers, between two non-preferential surfaces.

The effect of this confinement on the orientation of the copolymer morphology is striking. Eliminating preferential segmental affinities only at the substrate yields the atomic force microscopy image shown in Fig. 1a. The surface of this 137-nanometre-thick film is featureless because the lower surface energy of the polystyrene block means the lamellae orientate parallel to the air surface. Confining the block copolymer between two non-preferential surfaces, however, produces alignment of the lamellae normal to the surface of a 137-nanometre-thick film (Fig. 1b). This normal alignment of the lamellae covers the entire surface of the film and penetrates the whole film thickness, as demonstrated by neutron scattering and electron microscopy.

This strategy is directly applicable to any block copolymer system and to those having



Figure 1 Atomic force microscopic phase images. Images are from thin films of a P(S-b-MMA) symmetric diblock copolymer with a random copolymer (red) anchored either: **a**, to the substrate only; or **b**, to the substrate and air surface. Scale bar, 1 μm . Insets indicate the orientation of the lamellar morphology in the films.

a different morphology, such as cylinders. Our approach of controlling interfacial interactions allows a wide variety of nanoscopic structures to be produced, which potentially have applications as diverse as membrane separation⁷, quantum electronics⁸, nanolithography⁹ and catalysis¹⁰.

E. Huang*, L. Rockford*, T. P. Russell*, C. J. Hawker†

*Polymer Science and Engineering Department,
Silvio O. Conte National Center for Polymer
Research, University of Massachusetts,
Amherst, Massachusetts 01003, USA

†IBM Almaden Research Center,
650 Harry Road, San Jose,
California 95120-6099, USA

1. Anastasiadis, S. H., Russell, T. P., Satija, S. K. & Majkrzak, C. F. *Phys. Rev. Lett.* **62**, 1852–1855 (1989).
2. Mansky, P., Russell, T. P., Hawker, C. J., Pitsikalis, M. & Mayes, J. *Macromolecules* **30**, 6810–6813 (1997).
3. Pickett, G. T., Witten, T. A. & Nagel, S. R. *Macromolecules* **26**, 3194–3199 (1993).
4. Kikuchi, M. & Binder, K. *J. Chem. Phys.* **101**, 3367–3377 (1994).
5. Matsen, M. W. *J. Chem. Phys.* **106**, 7781–7791 (1997).
6. Mansky, P., Huang, E., Liu, Y., Russell, T. P. & Hawker, C. *Science* **275**, 1458–1460 (1997).
7. Martin, C. R. *Science* **266**, 1961–1965 (1994).
8. McEuen, P. L. *Science* **278**, 1729–1730 (1997).
9. Mansky, P., Chaikin, P. & Thomas, E. J. *Mater. Sci.* **30**, 1987–1992 (1995).
10. Mercier, L. & Pinnavaia, T. J. *Adv. Mater.* **9**, 500–503 (1997).

How to store seeds to conserve biodiversity

Zheng *et al.*¹ pointed out the potential advantages of ‘ultra-dry’ (less than 5% moisture content) seed storage for the long-term genetic conservation of plants, based largely on their studies of seed vigour. Here we support their suggestion, using our data from several years ago which showed that ultra-dried seeds have increased survival periods².

It has been shown that seeds of many species can benefit from being cooled and dried. Their longevity over a range of conditions can be described by an improved seed viability equation³, although there are limits to the range of conditions over which this equation applies. For long-term

seed storage, it is the lower temperature and moisture-content limits that are of most concern. The lower temperature limit is probably about $-20\text{ }^{\circ}\text{C}$ (ref. 3), although experiments at high temperatures (necessary to reduce longevity at low moisture contents to only a year or so) show that the low moisture-content limit is that produced by equilibrating seed at 10–12% relative humidity at $20\text{ }^{\circ}\text{C}$ before storage⁴. It is this protocol, combined with storage at ambient or cooler temperatures, that has been described as ‘ultra-dry’ seed storage⁵.

Although practical tests have indicated the benefits of ultra-dry storage^{6,7}, some controversy remains. It has been reported that ultra-dry storage may adversely affect seed vigour at certain temperatures⁸, although other results have not identified this problem⁹.

The results of Zheng *et al.*¹ support the practical application of ultra-dry seed storage as a low-cost, simple technology for gene banks. However, we question their suggestion¹ that this treatment provides longevity equal to that of more conventional gene banks. In some species at least, ultra-dry seed storage at ambient temperatures results in more rapid loss of seed viability than storage at about 5% moisture content at $-18\text{ }^{\circ}\text{C}$ (ref. 6).

Richard H. Ellis, Eric H. Roberts

Seed Science Laboratory,
Department of Agriculture,
The University of Reading,
PO Box 236, Reading RG6 6AT, UK
e-mail: r.h.ellis@reading.ac.uk

1. Zheng, G.-H., Jing, X.-M. & Tao, K.-L. *Nature* **393**, 223–224 (1998).
2. Ellis, R. H., Hong, T. D. & Roberts, E. H. *Ann. Bot.* **57**, 499–503 (1986).
3. Ellis, R. H. & Roberts, E. H. *Ann. Bot.* **45**, 13–30 (1980).
4. Ellis, R. H., Hong, T. D. & Roberts, E. H. *Ann. Bot.* **69**, 53–58 (1992).
5. *Annual Report 1991* (International Board for Plant Genetic Resources, Rome, 1992).
6. Ellis, R. H., Hong, T. D., Astley, D., Pinnegar, A. E. & Kraak, H. L. *Seed Sci. Technol.* **24**, 347–358 (1996).
7. Steiner, A. M. & Ruckebauer, P. *Seed Sci. Res.* **5**, 195–199 (1995).
8. Vertucci, C. W. & Roos, E. E. *Plant Physiol.* **94**, 1019–1023 (1990).
9. Ellis, R. H., Hong, T. D. & Roberts, E. H. *Ann. Bot.* **76**, 521–534 (1995).

Refrigeration can save seeds economically

Seed genebanks around the world are seeking economical ways to store seeds as a means of conserving plant biodiversity. Zheng *et al.*¹ suggest that the use of ‘ultra-dry’ technology^{2,3}, in which the seeds are dried to a water content of less than 5%, can extend the longevity of some seed species sufficiently to reduce or eliminate the need

for refrigeration. This would benefit in particular some developing countries, such as China, for which they say the cost of cold storage is prohibitive.

It is widely accepted that drying seeds can increase their longevity^{3–5}, and there is general agreement that, moisture contents being equal, the same seed lot will survive longer if stored at lower temperatures^{3–5}. The salient issue is whether the longevity achievable with various preservation protocols is sufficient for the conservation of germ plasm.

Base collections need seeds to survive for decades³. Zheng *et al.*¹ report good survival of ultra-dry *Eucommia* seeds after two years of storage at ambient temperatures. However, ultra-dry technology has several technical problems, including: (1) a general decline in germination within five years; (2) a narrow range of allowable moisture contents for seeds stored at high temperatures; and (3) a tremendous variability in longevity among different seed cultivars⁵.

Relatively successful storage has been reported for ultra-dry seeds of a single rape-seed cultivar in which 100% and 42% of seeds germinated after 11 and 18 years, respectively, but this is not representative of the performance of *Brassica* species under ultra-dry storage conditions⁵. Our germination records (www.ars-grin.gov/nssl; data available on request) show that refrigerated storage is required to achieve the seed longevity required for *ex situ* germplasm conservation.

Ultra-dry technology might seem to be economical, but this does not take into account the high cost of establishing the optimum water content for individual seed accessions or the labour-intensive expense of frequent viability monitoring and regenerations. Added to this is the risk of losing valuable germ plasm when seeds die during storage or accessions are regenerated from small populations. When these factors are considered, refrigeration, by comparison, is a bargain.

**Christina Walters, Eric E. Roos,
Darren H. Touchell, Phillip C. Stanwood,
Leigh Towill, Loren Wiesner,
Steve A. Eberhart**

USDA-Agricultural Research Service,
National Seed Storage Laboratory,
Fort Collins, Colorado 80523, USA
e-mail: chriv@lamar.colostate.edu

1. Zheng, G.-H., Jing, X.-M. & Tao, K.-L. *Nature* **393**, 223–224 (1998).
2. *Cost-effective Long-term Seed Stores* (International Board for Plant Genetic Resources, Rome, 1986).
3. *Report of the Expert Consultation on Genebank Standards* (Food and Agricultural Organization of the United Nations/International Board for Plant Genetic Resources, Rome, 1992).
4. Justice, O. L. & Bass, L. N. in *Principles and Practices of Seed Storage* (Agriculture Handbook no. 506, US Government Printing Office, Washington DC, 1978).
5. *Seed Sci. Res.* (suppl.) **8**, (1998).

Through the looking-glass

Mirror Image: Jonathan Miller on Reflection
Exhibition at the National Gallery, London,
until 13 December

On Reflection
by Jonathan Miller
National Gallery Publications: 1998. 224 pp.
£25, \$40

Jonathan Miller on Reflection
BBC TV series

Keith Miller

Hamlet tells his players that the purpose of art is to hold a mirror up to nature. The principle has a long ancestry, and has at times been interpreted surprisingly narrowly. In the fifth century BC, Zeuxis of Heraclea painted a still-life of grapes so lifelike that hungry birds flew down to peck at them, then was himself taken in by a curtain which seemed to cover a work by his rival Parrhasius, but turned out to be the work itself. The yardstick of artistic success in this and similar parables, repeated approvingly during the early Renaissance, was verisimilitude. The more completely a picture tricked you into thinking it was not a representation of a thing, but the thing itself, the better it was.

If art mirrors nature then an exhibition about mirrors in art ought to address some central questions. Jonathan Miller's show at the National Gallery, reflected, as it were, in a book, a short TV series and a panoply of shiny souvenirs, attempts to explain what reflection is, how it works and what it means in art. The results are uneven. Clearly, the exhibition intends to provide a sort of Alice-in-Wonderland experience. Miller's prolific text, displayed in various shapes and sizes, alternates with a mirrored corridor, a two-way mirror, a hall of mirrors and, for good measure, a few pictures. There is an undeniable jauntiness about all this, but inevitably the theatricality of the show muffles Miller's arguments, and, more problematically, tends to shout down the pictures, most of which are very interesting, and some of which are very beautiful.

Walking round the show is less an adventure through the looking-glass than a progress through the bowels of some grotesquely enlarged Dorling Kindersley textbook. A particular problem is Miller's promiscuous use of large-scale reproductions, many of them bad, of works that could not be obtained in person. This might touch on profound philosophical issues about truth and representation, but here it seems tatty and lazy. It is hard to avoid the conclusion that what Miller wants is to give us a talk with slides (in fact the genesis of the project was an invitation to do a couple of lectures).

The book that accompanies the exhibition



"Las Meninas" by Velázquez: a mirror on the experience of the artist himself.

is less a catalogue than an illustrated fleshing-out of the show's textual elements. Here the effect is less flashy and more lucid, although some of Miller's chatty ejaculations would discredit a youth TV anchorperson; egregious examples include "Hmmm . . .", and the even more appalling "(!)". The substance of the book is uncentred and somewhat repetitive, moving from the mechanics of reflection, to the process by which it is perceived, to some of the conventional meanings and symbolism of mirrors in art (ranging from Marian purity to vanity and self-absorption).

Some of Miller's observations are on the banal side, some seem blindingly obvious, many might well be interesting if one were to do a whole book about them. In fact someone has. Richard Gregory, whose *Mirrors in Mind* (reviewed in *Nature* 385, 785; 1997) deals with the physiology and psychology of how we see reflection, is an old friend of Miller's, and hopefully remains so despite having been so, let us charitably say, influential on Miller's project. In fact it is at times hard to see what

Miller has brought to the party. As good as he is on specifics (his explanation of why the eyes in some portraits seem to follow you around the room is exemplary, and shows that he realizes people want to know that sort of thing), the book is neither particularly accessible nor particularly substantial.

The best incarnation of *On Reflection* is probably Miller's quartet of little films, and it is here that his talents as a TV pundit are, naturally, most at home. He is concise and enthusiastic; his brio, aquiline gaze and exact timing animate the programmes impressively. Miller is clearly in love with the sound of his own voice — it is entirely appropriate that the mythical figure of Narcissus figures prominently in his argument. The programmes leave one wanting to look further, and know more.

Art critics talking about the project have tended to invoke the "two cultures" argument, which fingers Miller as a man of science, who has upset the humanist tribe by assuming that rigour and seriousness are his

exclusive birthright, while remaining blind to the sacred mysteries of artistic creation. But Miller has never exactly set the world of medicine on fire. What he is famous for these days is directing operas — not a profession in which scientific discipline is much of a *sine qua non* — and being on TV. The scientific element in *On Reflection* is not excessively taxing (try reading Michael Baxandall's book about shadows in eighteenth-century painting, *Shadow and Enlightenment*, Yale University Press, 1995). But throughout the project he shows a disappointing narrowness of response to the art works he discusses, seeming to think that a picture is only ever about one thing, taking no account of the childish pleasure that even the best artists sometimes take in pushing paint around, disregarding art-historical factors that are thumpingly relevant to certain works.

From this show you would never, for example, know that Parmigianino's self-portrait in a convex mirror reflects not only the difficulty of making flat ones at the time, but also that various sixteenth-century artists, for various reasons, reacted against the Renaissance ideal of verisimilitude in painting; nor that Tennyson's "Lady of Shalott" was a sort of leitmotiv among the later Romantics, for whom the image of someone condemned to see life through a mirror symbolized the tragic unreality of experience mediated by art. There are also basic errors of observation; you cannot see much of yourself if you stand in a pool, as Miller says Rembrandt's girlfriend can, partly because the water would be disturbed by your presence, and partly because your reflection would be grotesquely foreshortened. In any case, Hendrije is plainly not looking straight downwards in the picture; the reflective surface confers a tenderly erotic mood on the picture, but not in any strictly literal way.

Moreover, the pivotal importance of the mirror as a kind of private symbol among artists of the goals and limitations of what they are striving to do is largely overlooked. A dingy print of Velázquez's "Las Meninas", a picture of what the king and queen of Spain saw while he worked on their portrait in around 1656, and one of the most fascinating pictures about seeing and representing things ever painted, is in the show. Miller notices that the mirror on the back wall reflecting the royal couple is made to seem shiny and scuffed with incident white light — that, in his terms, it is an imperfect reflection. But the shininess of the mirror makes us aware that it is surrounded by pictures, which seem grimy and indistinct, but which can be seen to include paintings by Rubens of mortals punished for claiming that their skill in the arts rivalled the gods (an interesting foil to the bravura self-portrait in the foreground). Velázquez is often cited as an example of a court painter whose life was effectively one of imprisonment (one of his closest friends at court was the king's dwarf).

Among the many levels on which "Las Meninas" works, one is surely a gloriously brilliant analogy between the gifted but trapped artist and the mirror, which gives a perfect image but has, as it were, no say in the matter, being a passive redirector of whatever light an object throws at it. The fact that the mirror represents the imperious gaze of Velázquez's masters reinforces this sense of his predicament; the better he does his job, the more literal and reductive the result. It is surely this desire on the part of artists to do more than compete with the dumb mirror which would lead to their break from representation at the start of this century. □

Keith Miller is at 199d Camberwell Grove, London SE5 8JU, UK.

What two legs can learn from four legs

Foundations of Social Evolution

by Steven A. Frank

Princeton University Press: 1998. 268 pp.

\$49.50, £35 (hbk), \$16.95, £13.95 (pbk)

Cooperation Among Animals

by A. L. Dugatkin

Oxford University Press: 1998. 221 pp.

£45, \$65

Martin A. Nowak and Karl Sigmund

From Aristotle's *Zoon Politicon* to George Orwell's *Animal Farm*, philosophical and literary discussions of human social life have often alluded to its beastly side. Conversely, research on animal societies has brought science close to areas traditionally reserved for philosophy and the humanities — too close for comfort, some seem to think. It was all very well to derive convenient morals from animal fables. But a vocal in-group of biologists, a few of them best-seller writers, claim that our moral sense and our civic virtues are actually inherited from animals. These two books by A. L. Dugatkin and Steven Frank studiously avoid the minefield of human social evolution, but draw a considerable amount of their appeal from the resonances animal societies have with our own.

This holds particularly for issues of altruism and mutual help. Darwin saw in animal cooperation a "special difficulty" for his theory. Subsequent cohorts of evolutionary biologists have blown, in turn, hot and cold about the prevalence of selfishness in nature. The last generation, which set out to take the altruism out of altruism, was guided by two shining beacons, William D. Hamilton's "Genetical evolution of social behaviour" (*Journ. Theor. Biol.* **7**, 1–52; 1964) and Robert Trivers's "Evolution of reciprocal altruism" (*Quart. Rev. Biol.* **46**, 35–57; 1971). The former was based on kin selection: genes for helping relatives will spread, since relatives are diluted copies of oneself. The latter was



Pulling together: Asian weaver ants, *Oecophylla smaragdina*, cooperate in nest building.

based on reciprocity: help your helper and you help yourself. Both approaches were never meant to exclude other possibilities, and are, in particular, mutually compatible. Nevertheless, most of the subsequent contributions can be placed in one tradition or the other. This is the case with the two books at hand, despite the fact that both propose a synthesis. Dugatkin follows Trivers, Frank follows Hamilton, and they synthesize in different ways.

The books complement each other in another respect. Frank's work on the foundations of social evolution explores issues of method in a highly original, almost idiosyncratic manner, whereas Dugatkin's book opens a wide perspective on the sizeable literature dealing with field observations and lab experiments on animal cooperation.

Dugatkin manages to place more than 1,000 papers in context without getting flustered or entangled. This is due to the book's clear organization: there are separate chapters on cooperation in fishes, in birds, in non-primate mammals, in non-human primates and (no, not in humans) in social insects. In each chapter, there are sections devoted to different types of activities: grooming, foraging, anti-predator behaviour, and so on. And, in each section, the data are dissected according to the "four pillars of co-operation", categories of explanation described by Dugatkin in a clear and brisk theoretical introduction. These are, in addition to kin selection and reciprocal altruism, trait group selection and byproduct mutualism.

Group selection (in D. S. Wilson's sense) is not meant to imply a 'good for the species' type of argument, but is based on the benefits that devotion to a clique can confer to individuals. Dugatkin's treatment is fair and sympathetic, avoiding the confusing megasell so rampant with this issue. Byproduct

mutualism holds when the best one can do for oneself also happens to be best for the other. Most theoreticians fail to be excited by this type of cooperation which can obviously not be threatened by cheating, and byproduct mutualism has been branded a solution without a problem. But the concept may account for most instances of cooperation, as emerges from Dugatkin's empirical chapters.

Based on joint work with M. Mesterton-Gibbons (*Quart. Rev. Biol.* **67**, 267–281; 1992), Dugatkin describes all four categories in terms of a single game, the Cooperator's Dilemma, which is an extension of the notorious Prisoner's Dilemma. We are not sure whether this provides the neatest classification of the many facets of cooperation. But it certainly helps Dugatkin to focus a large part of his book on one crucial aspect — the rank order of the payoff values, which are measured in reproductive success. Empirical results on this issue are hard to get, and even in the best-studied case — predator inspection in fish — consensus has not been achieved.

Dugatkin gives an excellent description of this debate, which originated in an ingenious mirror experiment by Manfred Milinski (*Nature* **325**, 433–435; 1987), and of similarly contentious issues. But he shies away from entering the thicket of data and speculation on human cooperation, explaining his reluctance by the huge amount of work which has been done in this area by experimentalists with no evolutionary background. We think that his book might change this state of affairs, and suggest it as required reading in psychology classes.

Frank's book centres not on data, but on theories and even more on methods. It views natural selection as fitness maximization, and mostly deals with how to keep count of fitness, that is, contributions to the future of the population. This is to a large extent a question of book-keeping — how best to invest in future offspring, how to include the contributions of relatives, and so on — and it can lead, depending on social context and interactions, to real brain-teasers. Frank's treatment is based on his dexterous use of the Price equation, which states that the change in the average value of a character is proportional to the covariance of character and fitness (*Nature* **227**, 520–521; 1970). In the hands of Price (a brilliant and tragic figure who killed himself in 1974), and later in those of Hamilton and Franks, the covariance formula worked like magic, shedding light on such diverse topics as R. A. Fisher's fundamental theorem of natural selection, group selection arguments, or Hamilton's concepts of inclusive fitness and population viscosity.

Frank shows, for instance, that relatedness can be used as a measure both of genealogical kinship and of information about social partners. His applications are

not confined to the usual topics of social evolution, but range from parasite virulence to cytoplasmic incompatibility. One third of the book deals with sex allocation, the division of resources between daughters and sons. It allows Frank to elaborate on his three 'exchange mechanisms': reproductive value, kin selection and marginal value. In spite of such examples, the book is heavily slanted towards theory. Its use of techniques from many fields of genetics, statistics and economics makes it a demanding read. But it is worth the effort. Frank's 'how to' approach will certainly shape future modelling of social interactions. Both books are essential reading for those interested in the evolution of cooperation. □

Martin A. Nowak is at the Institute for Advanced Study, Princeton, New Jersey 08540, USA. Karl Sigmund is at the Institut für Mathematik, Universität Wien, Strudlhofgasse 4, A-1090 Vienna, Austria.

Fungus not always a bogeyman

Magical Mushrooms,
Mischievous Molds

by George W. Hudler
Princeton University Press: 1998. 248 pp.
\$29.95, £22.95

David L. Hawksworth

In 1986, India's foremost mycologist C. V. Subramanian remarked that fungi touch on every facet of our life and progress on Earth. Unseen, misunderstood, or regarded with suspicion, we cannot escape their influence. George Hudler from Cornell University, New York state, explains why in this most readable book. The multifarious topics tackled range from the beneficial to the harmful, and are approached by eclectic examples, albeit with a US bias. He admits he had a dilemma about what to include; I would

share the problem. Stories about potato blight and the Irish famine, the cereal disease ergot and witches, reindeer and radioactive lichens, bread, wine, mycotoxins, mycoses, and devastating tree diseases are all here.

No mention is made of *Pneumocystis carini*, the fungal lung infection causing pneumonia in so many people with AIDS, but did you know that the singer Bob Dylan was critically ill with histoplasmosis? Medicinal products have an appropriately high profile, but what of Ronald Hare's hypothesis on the origin of Fleming's first penicillin-producing isolate? The spores may have originated on the floor below Fleming's laboratory, where Charles La Touche was working on cultures from damp Victorian London basements.

Companies active in pharmaceutical bio-prospecting today recognize that fungi are so poorly known that locally obtained isolates can be as bioactive as ones from exotic tropical sites. A fungus that produces cyclosporin (used as an immunosuppressive drug) turned up only a few miles from Ithaca, in Cornell's own backyard.

In Asian culture, some fungi are legendary for their health benefits. Notable examples are *Ganoderma lucidum* (reishi) and *Lentinula edodes* (shiitake), which give rise to antitumour activity by enhancing the immune system, and are credited with many other medicinal properties. Their metabolites are the subject of ongoing research. Shiitake is now readily available in Western supermarkets, together with about 30 other species, and fungi merit a heightened profile in our diets from both health and ecological perspectives. Largely grown on crop residues, mushrooms are underestimated as potential players in the sustainable use of resources. A lack of imaginative recipes has limited their profile among gastronomes, but this is being rectified by *Hope's Mushroom Cookbook* (Mad River Press, 1993) in the United States, and *Patricia's Mushroom Cookery*, to be published by MycoNova this autumn in the United Kingdom.



Feel the itch: several fungi infect human skin. This one, *Microsporum gypseum*, causes ringworm.

Hudler is cautious in his advice on collecting for the pot — a topic that is emotive among collectors and conservationists, on grounds other than safety. Now practical guidance is available in the United Kingdom from the *Code of Conduct for Mushroom Collectors* issued by English Nature in September.

Hallucinogenic mushrooms, and, in particular, the experiences of Gordon Wasson and Timothy Leary, are given a high profile. But more about the religious aspects of fungi could have been expected from the title, for example the involvement in early Christianity posited for them in John Allegro's books.

Hudler covers the various symbiotic and parasitic relationships of fungi with other plants, the diseases fungi can cause in insects, and wood decay. But I missed a sense of the magnitude of fungal diversity, its role in global ecology, and potential contribution to improvements in agricultural production. Biocontrol is discussed, especially with reference to the gypsy moth in North America whose caterpillars are attacked by *Entomophaga maimaiga*, but what of the pioneering work on whitefly control in UK glasshouses using *Verticillium lecanii*, and recent field trials of *Metarrhizium anisopliae* in locust control in Africa?

The negative image of taxonomy is perpetuated by the use of some obsolete names. And why were the *Cladonia* and *Boletus* figures not identified? Hudler sees molecular approaches as the taxonomic panacea, but I find that view short-sighted. Around 1,800 new fungi are described each year, and some 100 sequences are added annually; the molecular knowledge gap is increasing exponentially.

Fungi lend themselves to illustrations that captivate. The half-tones and line drawings are of variable quality, but supplemented by eight pages of colour plates, mainly of macromycetes; regrettably none have scales. Sadly he passed up the opportunity to provide an entrée into the amazing beauty revealed by modern microscopy of the microfungi responsible for so much of the magic and mischief of the title.

This book prompts reflection on the nature of a world without fungi. Imagine if trees had not been able to expand from tropical to temperate and boreal regions, if wood never decayed, forest productivity dropped, there were no penicillins or cephalosporins, no food for numerous insects and other invertebrates, no lichen cover on the ground in the tundra, no mushrooms or bread to eat, no alcoholic drinks, and no human or plant fungal diseases.

Mycology suffers as an orphan in botanical science. It is a major discipline submerged in largely foreign terrain — departments and courses dealing with organisms of a different kingdom — and starved of appropriate texts. Hudler's book, in the tradition of Ernest Large's *Advance of the Fungi* (Jonathan Cape,

1940), fulfils its claim "to remind us just how vulnerable we are to the oft-neglected fungus-world". It deserves to be read by biologists at large, and merits a place in Christmas stockings and on student reading lists. □

David L. Hawksworth is at MycoNova, 114

Finchley Lane, London NW4 1DG, UK.

Twisted strands

A History of Molecular Biology

by Michel Morange

Harvard University Press: 1998. 342 pp.

\$39.95, £24.95

DNA Pioneers and Their Legacy

by Ulf Lagerkvist

Yale University Press: 1998. 184 pp. \$20,

£15.95

Sydney Brenner

Although molecular biology is still a relatively young science, it has already inspired a considerable corpus of historical writing. First, there is Horace Judson's account of the early days of molecular biology, *The Eighth Day of Creation*, and Robert Olby's account of the work that preceded the discovery of the structure of DNA, *The Path to the Double Helix*. Then there are all the memoirs and autobiographies written by the main participants, beginning, of course, with James Watson's *The Double Helix*.

Historians are inclined to distrust the personal memoir, which, in their view, relies too much on memory and is seen only from one point of view. They heed the message of Ryunosuke Akutagawa's story *Rashomon* (1915; better known through the later film), where different witnesses testify about a rape and murder, each, including the dead man himself, giving a different account of the same events. The scientific literature itself is not a very useful source as, in historical terms, most papers are constructs, deconstructing the historical events themselves, to reassemble them in a document written in a standard style and shaped by whatever journal editors and their referees happen to think is scientifically appropriate at the time.

Michel Morange is a biochemist who has written a history of molecular biology that also includes a history of genetic engineering, taking up the story more or less where Judson leaves off. He also wants to place more emphasis on the role of biochemistry, and to give the French School (as he calls it) a greater role in the history of the subject than has been ascribed to it by British and US writers. He throws interesting light on why the subject developed so slowly in France. Apart from the obstacles posed by a gerontocracy largely of biochemists, there was the legacy of conservative French biology which never quite accepted Darwin and the theory of natural selection. This is curiously French; their Cartesian views of the world had no

place for a theory in which nature behaved as an Anglo-Saxon empiricist.

Although he accepted it intellectually, Jacques Monod remained uncomfortable with Darwinian theory. In his book, *Chance and Necessity*, which brought the new ideas of biology to the French public, Monod thought that nature should have a *project*, that is, he wanted the finished product to exist somewhere so that natural selection could achieve it. It was also the French who proposed replacing the term teleology with teleonomy — a wolf in sheep's clothing with a sheep in sheep's clothing. I am happy to say that neither are much used today.

Morange also wishes to press on historians of molecular biology the importance of the distinction made by François Jacob and Monod between structural and regulatory genes. The distinction is, of course, much older than the 1960s, except that nobody could do anything about it until all the basic concepts of molecular biology had been established. Morange's account of the history of genetic engineering and of the developments that followed shows how molecular biology went from a highly intellectualized subject to one that became technologized and began to interact with the world outside. The heroes of the past are replaced by the managers and bureaucrats of today.

Ulf Lagerkvist also believes that it is important to explain science and scientists to the public and he thinks this is best done through a combination of history and personal memoir. He, too, moves biochemistry to the centre of the stage, and his history starts far back in the history of chemistry. But in other ways his book is very different. Miescher, the discoverer of DNA, is given a central role in the story. I knew that in 1895 Edmund Wilson wrote in his book, *The Cell in Development and Heredity*, that chromatin and its main component, nuclein, was the physical carrier of heredity, but I did not know that Miescher rejected this and entertained the absurd idea that inheritance was mediated by enantiomeric forms of protein molecules.

The book culminates with an account of the life and work of Arthur Kornberg, whose work on DNA polymerase is much admired by the author. The book has a certain charm to it, and it is interesting to have another Rashomonic view of the history where Watson and Crick do not occupy all of the stage and where the names of Luria, Delbrück and other members of the phage school do not even appear in the index.

Lagerkvist, like others of his generation, laments the fact that the students of today are not interested in the background of their science. Perhaps only the old read history, because the past is more familiar; the young are interested only in creating the future. □

Sydney Brenner is at the Molecular Sciences Institute, 2168 Shattuck Avenue, Berkeley, California 94704, USA.

Leptin and the regulation of body weight in mammals

Jeffrey M. Friedman & Jeffrey L. Halaas

The assimilation, storage and use of energy from nutrients constitute a homeostatic system that is essential for life. In vertebrates, the ability to store sufficient quantities of energy-dense triglyceride in adipose tissue allows survival during the frequent periods of food deprivation encountered during evolution. However, the presence of excess adipose tissue can be maladaptive. A complex physiological system has evolved to regulate fuel stores and energy balance at an optimum level. Leptin, a hormone secreted by adipose tissue, and its receptor are integral components of this system. Leptin also signals nutritional status to several other physiological systems and modulates their function. Here we review the role of leptin in the control of body weight and its relevance to the pathogenesis of obesity.

There is intense pressure to be thin in late-twentieth-century 'western' societies. Nevertheless, obesity is a prevalent condition which is often stigmatized. In addition, actuarial data indicate that life expectancy is reduced when body-mass index (BMI, body mass in kg/square of the height in metres) is 20% or more above the ideal¹. The ideal BMI is the level at which life expectancy is maximal and a BMI of >28 is now considered to represent obesity. The health risk of obesity is largely a consequence of the morbidities of diabetes, hypertension and heart disease, whose incidence increases with BMI.

The belief that obesity is largely the result of a lack of willpower, though widely held, is unsatisfactory. Studies of twins, analyses of familial aggregation, adoption studies and animal models of obesity all indicate that obesity is the result of both genetic and environmental factors^{2,3}. Moreover, weight is stable in lean and obese individuals even though much of the population actively practises weight control. Dieting is not usually successful in long-term maintenance of reduced body weight and most reduced obese individuals eventually regain the lost weight⁴.

The relative stability of weight in individuals indicated that energy balance may be controlled by a feedback loop, which maintains constancy of total body energy stores. It has been proposed that signals reflecting nutritional state are sensed by the hypothalamus, which, in turn, modulates food intake and energy expenditure^{5,6}. The identification of these signals proved difficult and for many decades their existence was widely questioned.

The *ob* gene and leptin

Recessive mutations in the mouse *obese* (*ob*) and *diabetes* (*db*) genes result in obesity and diabetes in a syndrome resembling morbid human obesity^{3,7}. *ob/ob* and *db/db* mice have identical phenotypes, each weighing three times more than normal mice (even when fed the same diet) and exhibiting a fivefold increase in body fat content. Data from cross-circulation (parabiosis) experiments suggested that the *ob* gene encoded, or was responsible for the generation of, a circulating factor that regulated energy balance and that the *db* gene encoded the receptor for this factor³. However, these conclusions were viewed with scepticism by many and their confirmation awaited the identification of the *ob* and *db* genes.

The cloning and characterization of the *ob* gene showed that it encodes a hormone, leptin (leptos (Greek): thin), that is expressed in adipose tissue and, at lower levels, in gastric epithelium and placenta⁸⁻¹⁰. The *ob* gene was isolated by positional cloning and found to encode a new, 167-amino-acid secreted protein⁸. The *ob* transcript is mutant in both of the available strains of *ob/ob* mice. In *ob*^{2j} mice, a ~5-kilobase (kb) ETn transposon is inserted into the first intron of *ob*¹¹. This mutation results in the synthesis of hybrid RNAs in which the splice donor of the non-coding *ob* first exon is joined to splice acceptors in the transposon. Mature *ob* RNA is not

synthesized in this mutant. In the C57BL/6J *ob/ob*^{1j} mutant, a nonsense mutation results in the synthesis of a truncated protein that is apparently degraded in the adipocyte⁸. In this mutant, the levels of *ob* RNA were increased, indicating that the gene might be under feedback control.

Initial data indicated that leptin might be an afferent signal in a negative-feedback loop regulating the size of adipose tissue mass (Fig. 1). If true, leptin RNA should be expressed in the principal site of fat storage, adipocytes; leptin should circulate in plasma; plasma leptin levels should correlate with adipose tissue mass; and recombinant leptin should reduce body fat content when injected into *ob/ob* and wild-type, but not into *db/db*, mice.

Leptin RNA is expressed in adipocytes, as shown using *in situ* hybridization, cell fractionation and immunohistochemistry¹². Leptin circulates as a protein of relative molecular mass 16,000 (*M_r* 16K) in mouse and human plasma¹³. A 19K form has also been identified in extracts of rat stomach, although the molecular nature and functional significance of this form are not known¹⁹. The 16K moiety is not modified post-translationally, as the relative molecular mass of the native protein is identical to that predicted by the primary structure (without the signal sequence)¹⁴. The plasma levels of leptin are highly correlated with adipose tissue mass and fall in both humans and mice after weight loss¹⁵. Levels of leptin are increased in obese humans and in several genetic and environmentally induced forms of rodent obesity^{2,15}.

Administration of leptin by injection or, with greater potency, as a constant subcutaneous infusion results in a dose-dependent decrease in body weight at incremental increases of plasma leptin levels within the physiological range^{13,16-19}. In both *ob/ob* and wild-type mice, leptin-induced weight loss is restricted to adipose tissue with sparing of lean body mass¹³.

There appears to be both a short-term and a long-term system for controlling feeding behaviour and energy balance. Plasma glucose concentration, body temperature, plasma amino acids, cholecystokinin (CCK) and other hormones can all modulate meal patterns (hunger and satiety) and are components of the short-term system^{20,21}. Leptin does not increase significantly after a meal and does not, by itself, lead to the termination of a meal^{15,22}. Leptin appears to function largely within the long-term system and influences the quantity of food consumed relative to the amount of energy expended. However, leptin is depleted from the stomach of rats after a meal or administration of CCK, indicating that it could also function in the short-term system or locally in the gastrointestinal tract⁹.

A broader role for leptin

ob/ob mice show many of the abnormalities seen in starved animals, including decreased body temperature, hyperphagia, decreased

energy expenditure (including activity), decreased immune function, and infertility³. Leptin replacement corrects all of these abnormalities, implying that *ob* mice exist in a state of 'perceived starvation' and that the resulting biological response in the presence of food leads to obesity^{13,17,23,24}. The idea that decreased plasma leptin levels signal nutrient deprivation is supported by the observation that exogenous leptin attenuates the neuroendocrine responses to food restriction²⁵. Fasted wild-type mice receiving leptin continue to ovulate, whereas fasted controls given saline experience an ovulatory delay of several days. Leptin treatment blunts the changes in circulating thyroid hormone and corticosterone levels that are normally associated with food deprivation²⁵. Starvation is also associated with decreased immune function and leptin corrects these abnormalities²⁴. Leptin stimulates proliferation of CD4⁺ T cells and increases production of cytokines by T-helper-1 cells²⁴. These results indicate that leptin may also be a key link between nutritional state and the immune system.

Leptin is also important in regulating the onset of puberty. Extremely thin women often stop ovulating and abnormally thin adolescent women enter puberty later than their heavier counterparts, indicating that fat tissue may produce a signal that regulates reproduction. This factor may be leptin. Treatment of mice with leptin accelerates the maturation of the female reproductive tract and leads to an earlier onset of the oestrous cycle and reproductive capacity^{26,27}. In humans, a surge in plasma leptin concentration is seen in prepubertal males²⁸. Leptin RNA and the leptin receptor have been identified in mouse and human placenta, but the functional significance of this is unknown^{10,29}.

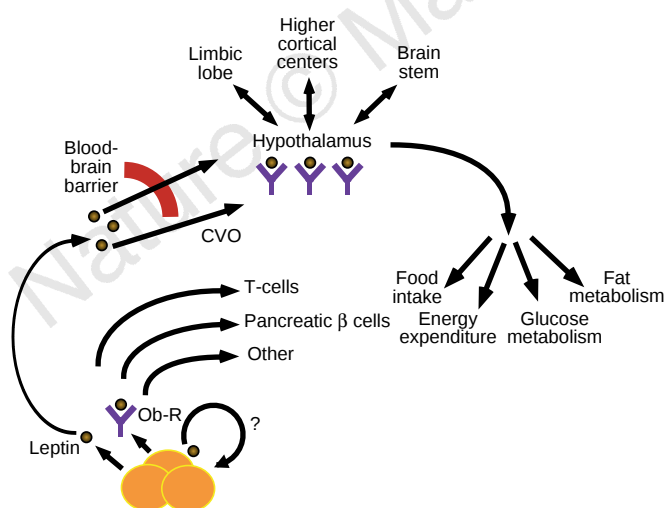


Figure 1 Adipocyte leptin and the regulation of adipose tissue mass. The cloning of the *ob* gene and the characterization of its gene product, leptin, indicated that body fat content may be under homeostatic control. Available data indicate that leptin is the afferent signal in a negative feedback loop that maintains constancy of adipose tissue mass. Leptin is secreted from adipocytes (bottom left) either as a 16K protein or bound to a soluble form of its receptor (Ob-R). The level of leptin is positively correlated with differences in body fat. Increased leptin results in negative energy balance (energy expenditure > food intake), whereas decreased levels lead to positive energy balance (food intake > energy expenditure). Leptin acts mainly on the hypothalamus (top). Extensive connections exist between the hypothalamus and other brain regions. Leptin acts centrally to decrease food intake and modulate glucose and fat metabolism (right). Peripheral effects on T cells, pancreatic islets and other tissues have also been demonstrated. CVO, circumventricular organ.

These observations led to speculation that leptin's main physiological role is to signal nutritional status during periods of food deprivation²¹. However, leptin's role in preventing excess weight gain has been shown to be physiologically significant. Lean mice given chronic infusions of leptin lose adipose mass in a dose-dependent manner at leptin levels within the physiological range¹⁶. Therefore, dynamic changes in plasma leptin concentration act to resist weight change in either direction.

Leptin receptor and leptin's sites of action

The leptin receptor (Ob-R) was first isolated from mouse choroid plexus by expression cloning³⁰. It was identified as a member of the cytokine family of receptors and binds leptin with nanomolar affinity³⁰. The stoichiometry of binding is unknown, but the Ob-R sequence probably includes two ligand-binding domains. Positional cloning of *db*, the Ob-R gene, showed that this gene encodes five alternatively spliced forms, Ob-Ra, Ob-Rb, Ob-Rc, Ob-Rd and Ob-Re (refs 30, 31). Ob-Rb (also known as Ob-RL) has a long cytoplasmic region containing several motifs required for signal transduction. The other forms lack some or all of these motifs (Fig. 2).

Mutations in Ob-R result in an obese phenotype identical to that of *ob* mice³². C57Bl/Ks *db/db* mice are phenotypically the same as other strains of *db* mice. In this co-isogenic strain, the Ob-Rb transcript contains an insert with a premature stop codon, as a result of abnormal splicing^{31,33}. Expression of Ob-Ra and the other splice forms is unaffected in this mutant. Thus Ob-Rb is essential for leptin's weight-reducing effects.

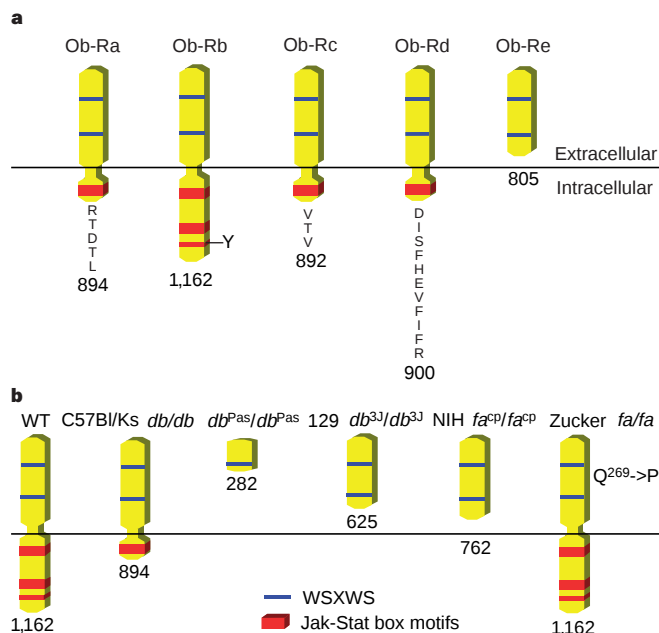


Figure 2 Leptin-receptor isoforms and receptor mutations in rodent models of obesity. **a**, There are at least five different isoforms of the leptin receptor in mouse. All share identical extracellular, ligand-binding domains but they differ at the C terminus. Four of the five have transmembrane domains, but only Ob-Rb encodes all protein motifs capable of activating the Jak-Stat signal transduction pathway. The remaining isoform, Ob-Re, is truncated before the membrane-spanning domain and is secreted. **b**, Mutations in Ob-R lead to massive obesity in *db* mice and *fa* rats. Most of the mutations affect all of the splice forms. However, in obese C57Bl/Ks *db/db* mice, the mutation creates a new splice donor that inserts a premature stop codon into the Ob-Rb 3'-end, resulting in the replacement of the Ob-Rb isoform by the Ob-Ra isoform. This mutation establishes the Ob-Rb isoform as being critical for leptin function. WT, wild type.

Although the Ob-Ra isoform, which has a short cytoplasmic domain, is expressed in the choroid plexus and many other tissues, its significance is unknown. Ob-Ra can activate gene expression and signal transduction in cultured cells, albeit weakly³⁴. It is unknown whether this occurs *in vivo*. The function of the other forms is likewise unclear³². They may function in the transport of leptin across the blood–brain barrier or form heterodimers with other cell-surface proteins.

Ob-Rb is normally expressed at high levels in hypothalamic neurons and in other cell types, including T cells and vascular endothelial cells^{24,31,35,36}. *In situ* hybridization was used to identify the hypothalamic arcuate nucleus, dorsomedial hypothalamic nucleus (DMH), paraventricular nucleus (PVN), ventromedial hypothalamic nucleus (VMH) and lateral hypothalamic nucleus (LH) as principal sites of Ob-Rb expression in the central nervous system (CNS). Each of these nuclei is important in regulating body weight^{5,37,38}. Leptin induces a dose-dependent activation of the transcription factor Stat3 in the hypothalamus of mice within 15 minutes of a single intraperitoneal injection³⁹. Stat3 was not activated in any other tissues. Leptin also increases the expression of *fos*, a Stat3 target, and several other genes in the hypothalamus specifically⁴⁰.

Synaptic transmission from neurons in the arcuate nucleus is altered by leptin⁴¹, which also leads to the hyperpolarization of some hypothalamic neurons and replicates the effect of glucose on these cells⁴². This effect depends on an ATP-dependent potassium channel, as tolbutamide, which blocks these channels, inhibits the effect⁴². These electrophysiological effects are rapid and unlikely to require new transcription through activation of STAT proteins. Although the components of leptin-mediated signal-transduction pathways that alter electrical activity in neurons are not yet known, their identification could provide new therapeutic targets.

Taken together with the fact that *db* mice and mice with hypothalamic lesions are leptin-resistant, these findings support the conclusion that the hypothalamus is an important site of leptin action^{12,13,43}. This conclusion is further supported by the observation that a single intracerebroventricular (i.c.v.) injection of leptin reduces food intake at doses that have no effect when delivered peripherally^{16,18,19}. Chronic i.c.v. infusion of leptin at a dose of 3 ng per hour results in total depletion of body adipose stores, whereas peripheral administration requires doses >500 ng h⁻¹ to achieve the same effect¹⁶. The effects of leptin on energy metabolism are the same regardless of the route of administration (refs 16 and 44, and our unpublished observations).

Low i.c.v. doses of leptin lead to a transient reduction in food intake that persists only until the adipose tissue mass has been depleted, after which food intake returns to normal¹⁶. The attenuation of the leptin response may be explained by the presence of other, undiscovered, signals, perhaps from skeletal muscle. These results also indicate that the effects of recombinant leptin are qualitatively different from those seen after parabiosis in which lean mice receiving *db/db* (hyperleptinaemic) plasma become anorectic and die of apparent starvation. A factor(s) other than leptin may be required for lethality after parabiosis of wild-type mice to *db* mice.

The mechanisms of leptin transport into the CNS is unknown. As leptin uptake occurs in the capillary endothelium of mouse and human brain, active transport by Ob-Ra or other proteins has been suggested as a possible mechanism⁴⁵. Some Ob-Rb-positive neurons project to the median eminence, the site of hypothalamic circumventricular organs, and may be in front of the blood–brain barrier⁴⁶. Circumventricular organs have fenestrations in their capillaries that allow direct communication between the blood and specific CNS neurons.

Leptin also acts on peripheral cell types and has direct mitogenic effects on CD4⁺ human T cells²⁴. Leptin also affects endothelial cells directly and increases angiogenesis, although high doses are

required³⁶. Leptin modulates pancreatic β -cell function *in vivo*⁴⁷. Leptin also has direct effects on other cell types *in vitro*. The leptin receptor is widely expressed, although the Ob-Ra (short) form of the receptor predominates in many of these tissues^{31,35,38}. Although the potency of i.c.v. leptin indicates that direct peripheral effects are not required for weight loss, the full spectrum of its actions is not known. It is thus possible that leptin acts on many tissues and functions in other physiological systems. The use of the *Lox-cre* system to introduce tissue-specific mutations into Ob-R may be required to establish the relevance of non-CNS sites of leptin action *in vivo*.

The introduction of mutations into Ob-Rb *in vitro* has been used to identify several sequences in the carboxy-terminal domain that are required to activate signal transduction^{34,48}. Activation of the leptin receptor depends on phosphorylation of the kinase Jak2 after ligand binding to an Ob-Rb homodimer. Receptor homodimerization may or may not be ligand-dependent: a baculovirus-expressed receptor spontaneously forms a homodimer⁴⁹. *In vitro*, binding of leptin to its receptor leads to the activation of Stat1, 3 and 5, whereas leptin activates only Stat3 *in vivo*^{35,39,48}. Leptin also leads to tyrosine phosphorylation of SHP-2, a phosphotyrosine phosphatase, which decreases both the state of Jak2 phosphorylation and transcription of a leptin-inducible reporter gene (ref. 50, and our unpublished observations). SOC-3 may also be involved in leptin-mediated signal transduction and it, as well as SHP-2, may downregulate the response to leptin⁵¹.

Leptin and the neuronal circuit regulating weight

Leptin receptors have been found in several hypothalamic nuclei, including the arcuate nucleus, VMH, LH, DMH and PVN^{5,37,38}. The LH and VMH project both within and outside the hypothalamus and modulate activity of the parasympathetic and sympathetic nervous systems, respectively. The DMH also has inputs to the parasympathetic nervous system and may be involved in integrating information among the VMH, LH and PVN. The PVN controls secretion of peptides from both the posterior and anterior pituitary and projects to nuclei with sympathetic or parasympathetic efferents.

These hypothalamic nuclei express one or more neuropeptides and neurotransmitters that regulate food intake and/or body weight (Table 1). Genetic data indicate that neuropeptide Y (NPY) and one or more of its receptors act in the response to absent (and possibly low) leptin, whereas melanocyte-stimulating hormone (MSH), its receptor, the melanocortin-4 receptor, and possibly the agouti-related transcript (ART, also known as AGRP; another melanocortin-4 receptor) are required for the response to an increased plasma leptin concentration (Fig. 3)⁵². NPY is the most potent orexigenic agent known when administered intrathecally. NPY RNA is increased in *ob/ob* mice and decreases after leptin treatment¹⁹. An NPY-knockout attenuates the obesity and other features of *ob/ob* mice⁵³.

Table 1 Hypothalamic modulators of food intake

Increase food intake	Decrease food intake
NPY	CART
MCH	CKK
Galanin	CRH
Orexin a and b	α -MSH
Peptide YY	Insulin
Noradrenaline	GLP-1
(α 2 receptor)	Bombesin
	Urocortin
	Serotonin

Many factors regulate feeding behaviour. The neurotransmitters and neuropeptides known to regulate food intake are shown. Leptin probably modulates the activity some or all of these factors (and vice versa). A detailed understanding of the functional relationships among leptin and these (and other) neuropeptides and neurotransmitters will be necessary to determine the mechanisms regulating food intake and body weight. GLP-1, glucagon-like peptide-1.

Abnormal melanocortin signalling in yellow agouti (A^Y) or melanocortin-4-knockout mice leads to obesity and leptin resistance^{16,54,55}. A subset of neurons expresses both Ob-R and proopiomelanocortin (POMC), and leptin modulates POMC gene expression (POMC is the precursor of MSH)⁴⁶. Agonists of α -MSH and MSH decrease food intake and pretreatment of animals with an α -MSH antagonist blunts the anorectic effect of injected leptin⁵⁶. ART, an endogenous antagonist of melanocortin signalling, is also implicated in the regulation of weight, as transgenic mice overexpressing ART are markedly obese⁵⁷. Levels of messenger RNA encoding this hypothalamic peptide are increased eightfold in *ob/ob* mice⁵⁸.

Other neurotransmitters and neuropeptides also function in this homeostatic system (Table 1)²¹. Leptin may stimulate the action of the anorexigenic agents and antagonize the orexigenic effects of others. Cholecystokinin (CCK) potentiates the anorectic effect of leptin⁵⁹. In addition, CCK-8 decreases stores of leptin in the stomach⁹. CART (cocaine- and amphetamine-regulated transcript), a hypothalamic peptide, decreases food intake; anti-CART antibodies increase food intake and levels of its mRNA are increased in *ob/ob* mice⁶⁰. Bombesin reduces food intake and induced mutations of the bombesin-3 receptor result in mild obesity⁶¹. Insulin acts on the hypothalamus to decrease food intake⁶². Levels of melanin-concentrating hormone (MCH), a neuropeptide expressed in the lateral hypothalamus, are increased in *ob/ob* mice and injections of it increase food intake in mice⁶³. Orexin-a and orexin-b are also expressed in the lateral hypothalamus and increase food intake⁶⁴. Leptin increases levels of corticotropin-releasing hormone (CRH) mRNA in the PVN and stimulates release of CRH from perfusion

slices of both amygdala and the PVN^{65,66}. Delivery of CRH to the PVN results in reduced food intake and increased energy expenditure in lean and obese rats⁶⁷. Pretreatment with anti-CRH antibodies decreases the anorectic effect of a single i.c.v. dose of leptin⁶⁸.

High levels of glucocorticoids are seen in most strains of genetically obese mice and adrenalectomy and glucocorticoid antagonists blunt the obesity evident in *ob/ob*, *db/db* and other obese mice⁶⁹. Low-dose glucocorticoid replacement restores the obese phenotype, indicating that glucocorticoids have a permissive role in the development of the obese phenotype⁶⁹. It is unknown whether the requirement for glucocorticoids in the development of the full *ob/ob* phenotype depends on the suppression of CRH or on other known effects of glucocorticoids, or whether another mechanism is involved.

The functional relationship among these molecules will have therapeutic implications. Antagonists of NPY and its receptors are being developed at present⁷⁰. Serotonin, which reduces food intake, and noradrenaline, which is orexigenic, are targets for known weight-reducing agents such as fen-fen (dexfenfluramine and phentermine). A knockout of the HT-2c form of the serotonin receptor leads to a form of mild obesity⁷¹.

In summary, leptin concentrations are sensed by groups of neurons in the hypothalamus. During starvation, leptin levels fall, thus activating a behavioural, hormonal and metabolic response that is adaptive when food is unavailable. Weight gain increases plasma leptin concentration and elicits a different response, leading to a state of negative energy balance. It is not yet known whether the same (or different) neurons respond to increasing and decreasing

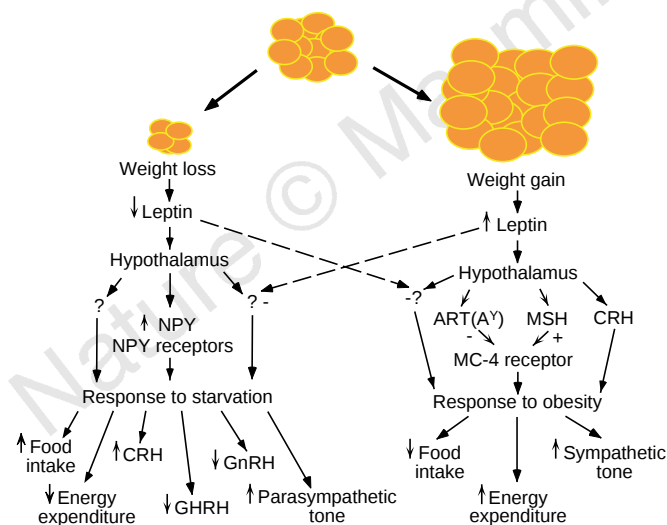


Figure 3 Biological response to high versus low leptin levels. Leptin acts as part of a feedback loop to maintain constant stores of fat. A loss of body fat (starvation) leads to a decrease in leptin, which in turn leads to a state of positive energy balance wherein food intake exceeds energy expenditure. A range of other metabolic and endocrine responses is also seen. Conversely, an increase in adiposity leads to an increase in the levels of leptin and a state of negative energy balance, with energy expenditure exceeding food intake. Genetic evidence indicates that different hypothalamic neuropeptides may mediate these responses. The melanocortin-4 (MC-4) receptor and its ligands, MSH and ART, are probably necessary for the biological response to increasing leptin levels. CRH also mediates some of leptin's effects, as pretreatment with an anti-CRH antibody blunts the anorectic effects of an i.c.v. dose of leptin. Other studies indicate that NPY is an important component of the biological response to low levels of leptin and possibly starvation. Other molecules are likely to play a role in the leptin response. These other factors are largely unknown but probably include some of the molecules shown in Table 1. GnRH, gonadotropin-releasing hormone; GHRH, growth-hormone-releasing hormone.

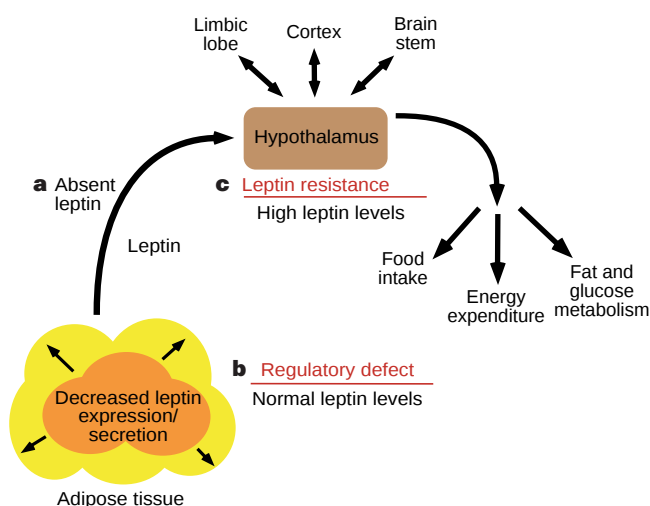


Figure 4 Pathogenesis of obesity. There are three general ways in which alterations of the leptin regulatory loop could lead to obesity. **a**, Failure to produce leptin, as occurs in *ob/ob* mice, would result in obesity, as would **b**, inappropriately low leptin secretion for a given fat mass. In the latter case, the fat mass would expand until 'normal' leptin levels are reached, resulting in obesity (also see Fig. 5). **c**, Finally, obesity could result from relative or absolute insensitivity to leptin at its site of action. Such resistance would be associated with increased circulating leptin, analogous to the increased insulin levels seen with insulin-resistant diabetes. In general, high plasma leptin levels are evident in obese rodents and humans. In a subset of cases, obesity is associated with normal levels of leptin. Differences in leptin production and leptin sensitivity could be the result of genetic, environmental and psychological factors.

leptin levels. The range of leptin's effects is likely to be complex, as different thresholds exist for several of leptin's actions⁷².

CNS efferent pathways regulating metabolism

The evidence that leptin acts centrally to reduce weight raises the question of how the CNS regulates peripheral metabolism in response to differences in leptin concentration. Increasing leptin leads to fatty acid oxidation and a reduction in adipose tissue mass, whereas leptin deficiency, seen in *ob/ob* mice or in mice receiving a leptin antagonist, is associated with an increase in fat deposition^{3,73}. It is unclear whether the metabolic derangements of *ob* mice are solely the result of increased food intake: other factors are likely to be important as food-restricted *ob* mice still deposit excess adipose tissue³.

Also, the mechanism by which centrally administered leptin leads to lipolysis and the loss of adipose tissue mass is unknown. The metabolic response to leptin is markedly different from the response to reduced food intake. Whereas food restriction (dieting) leads to the loss of both lean body mass and adipose tissue mass, leptin-induced weight loss is specific for the adipose tissue mass^{13,16}. Leptin also prevents the reduced energy expenditure normally associated with decreased food intake¹⁶. Finally, in contrast to food-restricted (pair-fed) animals, which show a marked increase in serum-free fatty acids, hyperleptinaemic animals undergoing a rapid period of weight loss fail to show a rise in serum-free fatty acids or ketones⁷⁴.

Leptin also has new effects on glucose metabolism. *ob/ob* mice are diabetic and the severity of the diabetes depends on the background strain carrying the mutation³. Leptin normalizes the hyperglycaemia and hyperinsulinaemia seen in C57BL/6J *ob/ob* mice at doses that do not decrease weight¹⁷. Treatment of lean animals with leptin leads to a reduction in serum glucose levels, without changing insulin levels, and increases glucose usage during euglycaemic clamp experiments⁷⁵. These effects are also seen after acute i.c.v. infusion of leptin, which leads to a significant increase in glucose turnover without altering the plasma insulin concentration⁴⁴. Previous data have indicated that the CNS has important effects on insulin secretion and glucose metabolism⁷⁶. If leptin increases glucose uptake in peripheral tissues in humans independently of weight loss, it could be of benefit to some patients suffering from non-insulin-dependent diabetes mellitus.

The mechanisms by which the CNS regulates fat and glucose metabolism in response to leptin are also unknown. Direct measurement of nerve activity has indicated that infusion of leptin increases sympathetic activity to brown adipose tissue, kidney, hind limb and adrenal gland⁷⁷. It is not yet known whether blockade of β -adrenergic receptors attenuates (or blocks) leptin-induced weight loss. Thus the role of the sympathetic nervous system in mediating the weight-reducing effect of leptin is not yet established.

An increase in sympathetic activity in brown-fat adipose tissue has also been observed in leptin-treated *ob/ob* mice⁷⁸. This is associated with an increase in the levels of uncoupling protein (UCP)-1 mRNA. Uncoupling proteins disrupt the mitochondrial proton gradient in brown fat (and possibly other tissues), resulting in the generation of heat rather than ATP. There is no evidence that leptin treatment of wild-type animals increases the activity of UCP-1. Levels of UCP-2 RNA may increase in leptin-treated mice, but the implications of this change are unknown⁷⁹. Leptin treatment does not cause a net increase in 24-hour energy expenditure but instead blunts the decreased energy expenditure that generally accompanies food restriction¹⁶. It is thus uncertain whether leptin increases energy expenditure or activates uncoupling protein.

Leptin and the pathogenesis of obesity

The role of leptin in the pathogenesis of obesity can be inferred by measurement of plasma leptin (Fig. 4). An increase in plasma leptin

suggests that obesity is the result of resistance to leptin. A low or normal plasma concentration of leptin in the context of obesity suggests decreased production of leptin. This interpretation is similar to that used in studies of insulin and the pathogenesis of type I and type II diabetes.

Plasma leptin has been measured in rodents and humans^{15,22}. Leptin circulates in a complex with a soluble form of Ob-R (Ob-Re) and it is not known whether assays of plasma leptin levels distinguish between bound and free leptin³². In general obese animals have higher leptin levels than controls (with the exception of *ob/ob* mice), indicating that these forms of animal obesity are associated with leptin resistance^{15,80}. Administration of leptin to obese hyperleptinaemic *A^y*, New Zealand obese (NZO) and diet-induced obese (DIO) mice confirmed that they are leptin-resistant^{16,81}. DIO mice are only partially leptin-resistant and lose moderate amounts of weight at high intraperitoneal doses of leptin. *A^y* mice (in which the response of the melanocortin-4 receptor to MSH is inhibited) are completely resistant to high i.c.v. doses of leptin, indicating that the normal function of the melanocortin-4 receptor is required for the response to exogenous leptin. NZO mice are resistant to peripherally administered leptin but respond normally to centrally administered leptin, suggesting that impaired leptin transport into the CNS causes this obesity¹⁶.

The cellular basis for leptin resistance in other obese animals is not known. Mutant *fat* and *tubby* mice are hyperleptinaemic¹⁵. Carboxypeptidase E (CPE), the gene product of the *fat* locus, alters post-translational processing of many peptides, including insulin, neurotensin, POMC and MCH^{82,83}. Mutations in PC-1, another peptide-processing enzyme, are also associated with obesity and increased leptin concentrations⁸⁴. The substrates of CPE and PC-1 that regulate weight and lead to obesity when these enzymes are defective are unknown. The *tub* gene product is highly expressed in the PVN^{85,86}. Further studies may reveal a link between this gene product and the response to leptin.

The leptin resistance seen in diet-induced obese AKR/J mice emphasizes the fact that environmental factors can modulate sensitivity to leptin^{16,81}. AKR/J mice remain lean when fed a standard chow diet but become obese when fed a high-fat diet. Other mouse strains do not become obese when fed an identical diet. This indicates that the pathogenesis of diet-induced obesity is the result of an interaction between genetic and environmental factors. An understanding of the mechanisms by which fat content in the diet modulates weight is likely to be relevant to human obesity as the incidence of obesity increases in many populations in proportion to the high-fat content of a 'western' diet. How might genes interact with environmental factors to cause obesity? A highly palatable diet often leads to transient weight gain. In most cases, the gained weight is rapidly lost. However, it is possible that in some cases, the induced increase in endogenous leptin levels (which accompanies weight gain) leads to a downregulation of the leptin response and a failure to return to the starting weight. If the downregulation of the leptin response is variable and influenced by genetic factors, one might predict that a subset of individuals would be particularly susceptible to diet-induced obesity. Alternative explanations are possible and further investigation is necessary. Identification of the AKR alleles that predispose AKR/J mice to diet-induced obesity may illuminate the underlying mechanisms⁸⁷.

The basis for leptin resistance in obese, hyperleptinaemic human subjects is unknown. Data from studies of animals indicate that this condition is likely to be heterogenous and that many factors may influence the activity of the neural circuit that regulates body weight. The entry of leptin into cerebrospinal fluid may be limiting in some obese subjects and morbid obesity could result when the plasma leptin levels exceed the capacity of the transport system^{88,89}. Factors that directly modulate energy expenditure or activate adipogenesis and lipogenesis could also result in apparent leptin

resistance. Finally, leptin's actions are likely to be influenced by psychological factors through connections between the higher cortical centres, which modulate an animal's motivational state, and the hypothalamus.

Regulation of leptin production

The physiological importance of quantitative changes in leptin concentration indicates that regulation of the *ob* gene is a critical control point^{16,25,72}. Decreased leptin expression per adipocyte could lead to obesity with normal (but inappropriately low) plasma leptin concentrations (Fig. 5). This prediction is supported by the observation that *ob/ob* mice carrying a poorly expressed leptin transgene are obese, despite having relatively normal leptin levels⁷².

The amount of *ob* gene expression per cell correlates with the lipid content and the corresponding size of individual adipocytes¹². The signal-transduction system by which fat cells sense their lipid content and adjust the level of *ob* expression is unknown. Studies of the regulation of *ob* expression have been hindered by the fact that cultured adipocyte cell lines deposit very little lipid relative to adipocytes *in vivo* and hence express only small amounts of leptin¹². The observation that cultured adipocytes express physiological levels of *ob* RNA when introduced directly into animals suggests an approach for studying quantitative changes in *ob* gene expression⁹⁰.

Extrinsic factors also modulate the expression of leptin. Leptin levels increase by ~30% at night²². Although feeding does not

appear to increase leptin expression, fasting acutely decreases plasma leptin concentrations^{15,22}. High intracellular glucosamine concentrations increase leptin production in adipose tissue and skeletal muscle⁹¹. Tumour-necrosis factor, insulin, glucocorticoids, interleukin-1 and other proteins also modulate *ob* gene expression^{92,93}. The *ob* promoter is responsive to several transcription factors, including C/EBP and peroxisome proliferator-activated receptor- γ ⁹²⁻⁹⁵. It is not known whether leptin is regulated and released from storage vesicles or secreted through a constitutive pathway.

Leptin in human physiology

Plasma leptin concentration correlates with body fat content and is usually increased in obese subjects^{15,22}. This suggests that human obesity is generally associated with an insensitivity to leptin. However, 5–10% of obese human subjects have relatively low levels of leptin, indicative of a reduced rate of leptin production in this subgroup^{15,22}. Low leptin levels also predispose pre-obese Pima Indians to weight gain⁹⁶.

In humans, diet-induced weight loss results in a decrease in plasma leptin concentration^{15,22}. This may explain the high failure rate of dieting, as low leptin is likely to be a potent stimulus to weight gain. Anorexia nervosa patients also have extremely low leptin levels^{97,98}. Refeeding of these patients results in a rapid increase in plasma leptin concentration to roughly normal levels before normal weight is achieved. Thus, excessive leptin production could play a permissive role in the pathogenesis of this condition.

In almost all cases, obese subjects express at least some leptin, indicating that human *ob* mutations are likely to be rare. Indeed there were no *ob* mutations in one study in which ~500 obese subjects were tested⁹⁹. Nevertheless, a few *ob* mutations have been described. Two cousins homozygous for a frameshift mutation in the leptin gene are markedly obese and do not have any circulating leptin¹⁰⁰. Three members of a Turkish kindred with a missense mutation in the leptin gene are extremely obese and amenorrhoeic, indicating that leptin is important in modulating human reproductive function¹⁰¹. Three massively obese members of a French family carry mutations in the leptin receptor and have reproductive abnormalities¹⁰². It is not yet known whether heterozygous members of these families show a more subtle phenotype. Hypercortisolaemia, cold intolerance and severe diabetes, abnormalities of *ob/ob* mice, are not seen in leptin-deficient humans.

The *ob* gene has been linked to severe obesity in some, but not all, family studies, but mutations in the leptin-coding sequence were not identified^{103,104}. The molecular basis for this linkage is unknown but could relate to differences in the amount of expression of leptin mRNA. Loci on human chromosome 2 may be linked to leptin levels and, to a lesser extent, to BMI¹⁰⁵. These loci are near to the human POMC gene, and two obese, red-haired subjects with POMC mutations have been identified¹⁰⁶.

The association of mutations in leptin and its receptor with massive obesity confirms its importance in regulating human body weight. However, these syndromes are rare. The pathogenesis of most human obesity is unknown and likely to be the result of differences in leptin secretion and/or leptin sensitivity. Both genetic and physiological studies are required to confirm this prediction.

Prospects for leptin treatment of human obesity

The possible therapeutic benefit of leptin treatment in humans is now being studied in clinical trials. Early data show that 4 weeks of daily leptin injections are safe and cause small but significant weight loss in lean and obese subjects, compared with placebo effects ($P < 0.02$) (A. S. Greenberg *et al.*, unpublished observations). A subset of eight obese subjects treated for a total of six months (0.3 mg kg^{-1} leptin subcutaneously) lost 7.1 kg compared with 1.7 kg in a group receiving placebo. Some of the subjects in this group lost substantial amounts of weight, but others did not. This

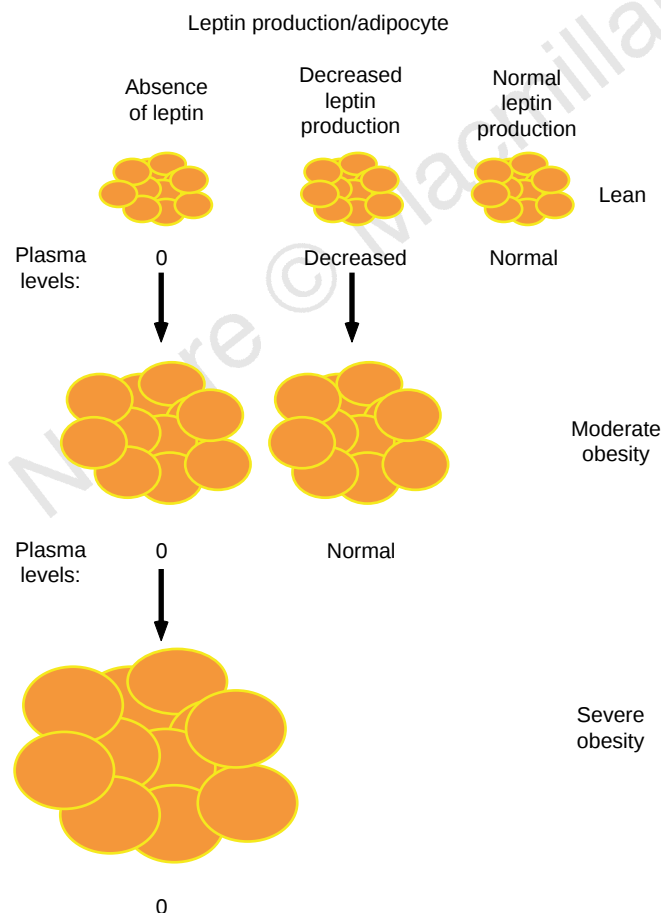


Figure 5 Regulation of leptin expression and the development of obesity. A relative decrease in leptin production by adipose tissue would be expected to lead to obesity. In such a case, the fat cell mass would be predicted to expand until the set point for plasma leptin levels was reached. Thus obesity with normal leptin levels is inferred to result from a relative decrease in leptin production by adipose tissue.

limited study indicates that leptin might be an effective therapy for some obese subjects, although more patients need to be investigated.

The therapeutic response has not yet been related to the initial plasma leptin level. Perhaps only the subset of obese individuals with low plasma leptin (5–10%) will be completely sensitive to exogenous leptin. This possibility is supported by the robust response to leptin treatment of obese transgenic mice that hypo-secrete leptin⁷². Leptin administration could also produce a biological response in those obese individuals who have high circulating levels of leptin, as is the case with insulin treatment of many type II diabetes patients. This may require the administration of relatively high doses.

The amount of protein that can be delivered is limited by its solubility and by local reactions in the skin that occur in response to high doses. Improvements in the leptin formulation and refinements in the dosing may mitigate these difficulties. Animal studies indicate that the ability to optimize the pharmacokinetics of recombinant leptin may greatly influence its usefulness. For example, leptin is more potent when administered as a subcutaneous infusion, whereas once-daily intraperitoneal injections are largely ineffective^{13,16–19}. Direct administration of leptin into cerebrospinal fluid is very effective but its use would require that the drug be delivered safely¹⁶. Gene therapy may provide an alternative means of delivering leptin. Adenovirus and adeno-associated virus vectors expressing leptin both exert potent weight-reducing effects in rodents^{74,107}.

Prospective epidemiological studies are required to define the indications for antiobesity treatments. The health risk of obesity is greatly diminished when even moderate amounts of weight (5% of total weight) are lost¹⁰⁸. This is the result of a marked improvement in the diabetic, hypertensive and cardiovascular status of obese subjects affected by these conditions¹⁰⁸. In rodents, leptin can increase glucose metabolism independently of effects on weight, an observation that indicates a possible use for it in the treatment of type II diabetes^{17,44,75}. The presence of co-morbidities is an obvious indication because dieting, by itself, is only rarely effective for the long-term maintenance of weight loss⁴. It is less clear whether leptin (or any other antiobesity treatment) is indicated in the absence of co-morbid conditions. This issue is complicated by the observation in one study that women weighing 15% less than average had reduced mortality rates, irrespective of co-morbidities¹⁰⁹.

Whether leptin finds its way into general usage as an antiobesity drug, the use of modern methods to identify and target the components of the leptin signalling pathway will form the basis for new pharmacological approaches to the treatment of obesity and other nutritional disorders. Further studies of leptin are also likely to reveal additional links between nutritional state and animal physiology.

J. M. Friedman is at the Laboratory of Molecular Genetics and Howard Hughes Medical Institute, and J. L. Halaas is at the Laboratory of Molecular Genetics, The Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.

- Harris, T. *et al.* Body mass index and mortality among nonsmoking older persons: the Framingham Heart Study. *J. Am. Med. Assoc.* **259**, 1520–1524 (1988).
- Stunkard, A. J., Harris, J. R., Pedersen, N. L. & McClearn, G. E. The body-mass index of twins who have been reared apart. *N. Engl. J. Med.* **322**, 1483–1487 (1990).
- Coleman, D. L. Obese and Diabetes: two mutant genes causing diabetes-obesity syndromes in mice. *Diabetologia* **14**, 141–148 (1978).
- Wadden, T. A. Treatment of obesity by moderate and severe caloric restriction. Results of clinical research trials. *Ann. Intern. Med.* **119**, 688–693 (1993).
- Hetherington, A. W. & Ranson, S. W. The spontaneous activity and food intake of rats with hypothalamic lesions. *Am. J. Physiol.* **136**, 609–617 (1942).
- Kennedy, G. C. The role of depot fat in the hypothalamic control of food intake in the rat. *Proc. R. Soc. Lond. B* **140**, 578–592 (1953).
- Friedman, J. M. & Leibel, R. L. Tackling a weighty problem. *Cell* **69**, 217–220 (1992).
- Zhang, Y. *et al.* Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**, 425–432 (1994).
- Bado, A. *et al.* The stomach is a source of leptin. *Nature* **394**, 790–793 (1998).
- Masuzaki, H. *et al.* Nonadipose tissue production of leptin: leptin as a novel placenta-derived hormone in humans. *Nature Med.* **3**, 1029–1033 (1997).
- Moon, B. C. & Friedman, J. M. The molecular basis of the obese mutation in *ob²* mice. *Genomics* **42**, 152–156 (1997).
- Maffei, M. *et al.* Increased expression in adipocytes of *ob* RNA in mice with lesions of the hypothalamus and with mutations at the *db* locus. *Proc. Natl Acad. Sci. USA* **92**, 6957–6960 (1995).
- Halaas, J. L. *et al.* Weight-reducing effects of the plasma protein encoded by the obese gene. *Science* **269**, 543–546 (1995).
- Cohen, S. L. *et al.* Characterization of endogenous human leptin. *Nature* **382**, 589 (1996).
- Maffei, M. *et al.* Leptin levels in human and rodent: measurement of plasma leptin and *ob* RNA in obese and weight-reduced subjects. *Nature Med.* **1**, 1155–1161 (1995).
- Halaas, J. L. *et al.* Physiological response to long-term peripheral and central leptin infusion in lean and obese mice. *Proc. Natl Acad. Sci. USA* **94**, 8878–8883 (1997).
- Pelleymounter, M. A. *et al.* Effects of the obese gene product on body weight regulation in *ob/ob* mice. *Science* **269**, 540–543 (1995).
- Campfield, L. A. *et al.* Recombinant mouse OB protein: evidence for a peripheral signal linking adiposity and central neural networks. *Science* **269**, 546–549 (1995).
- Stephens, T. W. *et al.* The role of neuropeptide Y in the antiobesity action of the obese gene product. *Nature* **377**, 530–534 (1995).
- Schneider, B. S., Friedman, J. M. & Hirsch, J. in *Brain Peptides* (eds Krieger, D., Martin, J. & Brownstein, M.) 251–280 (Wiley, USA, 1983).
- Spiegelman, B. M. & Flier, J. S. Adipogenesis and obesity: rounding out the big picture. *Cell* **87**, 377–389 (1996).
- Considine, R. V. *et al.* Serum immunoreactive-leptin concentrations in normal-weight and obese humans. *N. Engl. J. Med.* **334**, 324–325 (1996).
- Chehab, F. F., Lim, M. E. & Lu, R. Correction of the sterility defect in homozygous obese female mice by treatment with the human recombinant leptin. *Nature Genet.* **12**, 318–320 (1996).
- Lord, G. M. *et al.* Leptin modulates the T-cell immune response and reverses starvation induced immunosuppression. *Nature* **394**, 897–899 (1998).
- Ahima, R. S. *et al.* Role of leptin in the neuroendocrine response to fasting. *Nature* **382**, 250–252 (1996).
- Chehab, F. F., Mounzih, K., Lu, R. & Lim, M. E. Early onset of reproductive function in normal female mice treated with leptin. *Science* **275**, 88–90 (1997).
- Ahima, R. S. *et al.* Leptin accelerates the onset of puberty in normal female mice. *J. Clin. Invest.* **99**, 391–395 (1997).
- Mantzoros, C. S., Flier, J. S. & Rogol, A. D. A longitudinal assessment of hormonal and physical alterations during normal puberty in boys. Rising leptin levels may signal the onset of puberty. *J. Clin. Endocrinol. Metab.* **82**, 1066–1070 (1997).
- Gavrilou, O., Barr, V., Marcus-Samuels, B. & Reitman, M. Hyperleptinemia of pregnancy associated with the appearance of a circulating form of the leptin receptor. *J. Biol. Chem.* **272**, 30546–30551 (1997).
- Tartaglia, L. A. *et al.* Identification and expression cloning of a leptin receptor, OB-R. *Cell* **83**, 1263–1271 (1995).
- Lee, G. H. *et al.* Abnormal splicing of the leptin receptor in *diabetic* mice. *Nature* **379**, 632–635 (1996).
- Li, C. *et al.* Absence of soluble leptin receptor in plasma from *db^{fa}/db^{fa}* and other *db/db* mice. *J. Biol. Chem.* **273**, 10078–10082 (1998).
- Chen, H. *et al.* Evidence that the diabetes gene encodes the leptin receptor: identification of a mutation in the leptin receptor gene in *db/db* mice. *Cell* **84**, 491–495 (1996).
- Bjorbaek, C., Uotani, S., da Silva, B. & Flier, J. S. Divergent signaling capacities of the long and short isoforms of the leptin receptor. *J. Biol. Chem.* **272**, 32686–32695 (1997).
- Ghilardi, N. *et al.* Defective STAT signaling by the leptin receptor in *diabetic* mice. *Proc. Natl Acad. Sci. USA* **93**, 6231–6235 (1996).
- Sierra-Honigsmann, M. R. *et al.* Biologic action of leptin as an angiogenic factor. *Science* **281**, 1683–1685 (1998).
- Mercer, J. G. *et al.* Localization of leptin receptor mRNA and the long form splice variant (Ob-Rb) in mouse hypothalamus and adjacent brain regions by *in situ* hybridization. *FEBS Lett.* **387**, 113–116 (1996).
- Fei, H. *et al.* Anatomic localization of alternatively spliced leptin receptors (Ob-R) in mouse brain and other tissues. *Proc. Natl Acad. Sci. USA* **94**, 7001–7005 (1997).
- Vaisse, C. *et al.* Leptin activation of Stat3 in the hypothalamus of wild-type and *ob/ob* mice but not *db/db* mice. *Nature Genet.* **14**, 95–97 (1996).
- Woods, A. J. & Stock, M. J. Leptin activation in hypothalamus. *Nature* **381**, 745 (1996).
- Glaum, S. R. *et al.* Leptin, the obese gene product, rapidly modulates synaptic transmission in the hypothalamus. *Mol. Pharmacol.* **50**, 230–235 (1996).
- Spanwick, D. *et al.* Leptin inhibits hypothalamic neurons by activation of ATP-sensitive potassium channels. *Nature* **390**, 521–525 (1997).
- Satoh, N. *et al.* Pathophysiological significance of the obese gene product, leptin, in ventromedial hypothalamus (VMH)-lesioned rats: evidence for loss of its satiety effect in VMH-lesioned rats. *Endocrinology* **138**, 947–954 (1997).
- Kamohara, S., Burcelin, R., Halaas, J. L., Friedman, J. R. & Charron, M. J. Acute stimulation of glucose metabolism in mice by leptin treatment. *Nature* **389**, 374–377 (1997).
- Banks, W. A. *et al.* Leptin enters the brain by a saturable system independent of insulin. *Peptides* **17**, 305–311 (1996).
- Hakansson, M.-L. *et al.* Leptin receptor immunoreactivity in chemically defined target neurons of the hypothalamus. *J. Neurosci.* **18**, 559–572 (1998).
- Wang, M. Y. *et al.* Ob-Rb gene transfer to leptin-resistant islets reverses diabetogenic phenotype. *Proc. Natl Acad. Sci. USA* **95**, 714–718 (1998).
- White, D. W. *et al.* Leptin receptor (Ob-R) signaling. Cytoplasmic domain mutational analysis and evidence for receptor homo-oligomerization. *J. Biol. Chem.* **272**, 4065–4071 (1997).
- DeVos, R. *et al.* Ligand-independent dimerization of the extracellular domain of the leptin receptor and determination of the stoichiometry of leptin binding. *J. Biol. Chem.* **272**, 18304–18310 (1997).
- Carpenter, L. R. *et al.* Enhancing leptin response by preventing SH2-containing phosphatase 2 interaction with Ob receptor. *Proc. Natl Acad. Sci. USA* **95**, 6061–6066 (1998).
- Bjorbaek, C. *et al.* Identification of SOX-3 as a potential mediator of central leptin resistance. *Mol. Cell* **1**, 619–625 (1998).
- Friedman, J. M. The alphabet of weight control. *Nature* **385**, 119–120 (1997).
- Erickson, J. C., Hollopeter, G. & Palmiter, R. D. Attenuation of the obesity syndrome of *ob/ob* mice by the loss of neuropeptide Y. *Science* **274**, 1704–1707 (1996).
- Fan, W. *et al.* Role of melanocortinergic neurons in feeding and the agouti obesity syndrome. *Nature* **385**, 165–168 (1997).
- Huszar, D. *et al.* Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell* **88**, 131–141 (1997).
- Satoh, N. *et al.* Satiety effect and sympathetic activation of leptin are mediated by hypothalamic melanocortin system. *Neurosci. Lett.* **249**, 107–110 (1998).
- Ollmann, M. M. *et al.* Antagonism of central melanocortin receptors *in vitro* and *in vivo* by agouti-related protein. *Science* **278**, 135–138 (1997).

58. Shutter, J. R. *et al.* Hypothalamic expression of ART, a novel gene related to agouti, is up-regulated in obese and diabetic mutant mice. *Genes Dev.* **11**, 593–602 (1997).
59. Matson, C. A., Wiater, M. F., Kuijper, J. L. & Weigle, D. S. Synergy between leptin and cholecystokinin (CCK) to control daily caloric uptake. *Peptides* **18**, 1275–1278 (1997).
60. Kristensen, P. *et al.* Hypothalamic CART is a new anorectic peptide regulated by leptin. *Nature* **393**, 72–76 (1998).
61. Ohi-Hamazaki, H. *et al.* Mice lacking bombesin receptor subtype-3 develop metabolic defects and obesity. *Nature* **390**, 165–169 (1997).
62. Woods, S. C., Chavez, M. & Park, C. R. The evaluation of insulin as a metabolic signal influencing behavior via the brain. *Neurosci. Biobehav. Rev.* **20**, 139–144 (1996).
63. Qu, D. *et al.* A role for melanin-concentrating hormone in the central regulation of feeding behavior. *Nature* **380**, 243–247 (1996).
64. Sakurai, T. *et al.* Orexins and orexin receptors: A family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92**, 573–585 (1998).
65. Schwartz, M. W. *et al.* Identification of targets of leptin action in rat hypothalamus. *J. Clin. Invest.* **98**, 1101–1106 (1996).
66. Elmquist, J. K. *et al.* Leptin activates neurons in ventrobasal hypothalamus and brainstem. *Endocrinology* **138**, 839–842 (1997).
67. Rohner-Jeanrenaud, F., Walker, C.-L., Greco-Perotto, R. & Jeanrenaud, B. Central corticotropin-releasing factor administration prevents the excessive body weight gain of genetically obese (fa/fa) rats. *Endocrinology* **124**, 733–739 (1989).
68. Gardner, J. D., Rothwell, N. J. & Luheshi, G. N. Leptin affects food intake via CRF-receptor-mediated pathways. *Nature Neurosci.* **1**, 103 (1998).
69. Freedman, M. R., Horwitz, B. A. & Stern, J. S. Effect of adrenalectomy and glucocorticoid replacement on development of obesity. *Am. J. Physiol. Soc.* R595–R607 (1986).
70. Gerald, C. *et al.* A receptor subtype involved in neuropeptide-Y induced food intake. *Nature* **382**, 168–171 (1996).
71. Tecott, L. H. *et al.* Eating disorder and epilepsy in mice lacking 5-HT_{2c} serotonin receptors. *Nature* **374**, 542–546 (1995).
72. Ioffe, E., Moon, B., Connolly, E. & Friedman, J. M. Abnormal regulation of the leptin gene in the pathogenesis of obesity. *Proc. Natl Acad. Sci. USA* **95**, 11852–11857 (1998).
73. Verplogen, S. *et al.* A human leptin mutant induces weight gain in normal mice. *FEBS Lett.* **405**, 237–240 (1997).
74. Chen, G. *et al.* Disappearance of body fat in normal rats induced by adenovirus-mediated leptin gene therapy. *Proc. Natl Acad. Sci. USA* **93**, 14795–14799 (1996).
75. Sivitz, W. I. *et al.* Effects of leptin on insulin sensitivity in normal rats. *Endocrinology* **138**, 3395–3401 (1997).
76. Miles, P. D. G., Yamatani, K., Lickley, H. L. A. & Vranic, M. Mechanism of glucoregulatory responses to stress and their deficiency in diabetes. *Proc. Natl Acad. Sci. USA* **88**, 1296–1300 (1991).
77. Haynes, W. G. *et al.* Receptor-mediated regional sympathetic nerve activation by leptin. *J. Clin. Invest.* **100**, 270–278 (1997).
78. Collins, S. *et al.* Role of leptin in fat regulation. *Nature* **380**, 677 (1996).
79. Zhou, Y.-T. *et al.* Induction by leptin of uncoupling protein-2 and enzymes of fatty acid oxidation. *Proc. Natl Acad. Sci. USA* **94**, 6386–6390 (1997).
80. Frederich, R. C. *et al.* Leptin levels reflect body lipid content in mice: evidence for diet-induced resistance to leptin action. *Nature Med.* **1**, 1311–1314 (1995).
81. Van Heek, M. *et al.* Diet-induced obese mice develop peripheral, but not central, resistance to leptin. *J. Clin. Invest.* **99**, 385–390 (1997).
82. Naggert, J. K. *et al.* Hyperproinsulinaemia in obese fat/fat mice associated with a carboxypeptidase E mutation which reduces enzyme activity. *Nature Genet.* **10**, 135–142 (1995).
83. Rovere, C., Viale, A., Nahon, J. L. & Kitabgi, P. Impaired processing of brain proneurotensin and pro-melanin-concentrating hormone in obese fat/fat mice. *Endocrinology* **137**, 2954–2958 (1996).
84. Jackson, R. S. *et al.* Obesity and impaired prohormone processing associated with mutations in the human prohormone convertase 1 gene. *Nature Genet.* **16**, 303–306 (1997).
85. Kleyn, P. W. *et al.* Identification and characterization of the mouse obesity gene *tubby*: a member of a novel gene family. *Cell* **85**, 281–290 (1996).
86. Noben-Trauth, K., Naggert, J., North, M. & Nishina, P. A candidate for the mouse mutation *tubby*. *Nature* **380**, 534–538 (1996).
87. West, D. B., Boozer, C. N., Moody, D. L. & Atkinson, R. L. Dietary obesity in nine inbred mouse strains. *Am. J. Physiol.* **262**, R1025–R1032 (1992).
88. Schwartz, M. W. *et al.* Cerebrospinal fluid leptin levels: relationship to plasma levels and to adiposity in humans. *Nature Med.* **2**, 589–593 (1996).
89. Caro, J. F. *et al.* Decreased cerebrospinal-fluid/serum leptin ratio in obesity: a possible mechanism for leptin resistance. *Lancet* **348**, 159–161 (1996).
90. Mandrup, S. *et al.* Obese gene expression at *in vivo* levels by fat pads derived from Sc implanted 3T3-F442a preadipocytes. *Proc. Natl Acad. Sci. USA* **94**, 4300–4305 (1997).
91. Wang, J. *et al.* A nutrient-sensing pathway regulates leptin gene expression in muscle and fat. *Nature* **393**, 384–388 (1998).
92. Saladin, R. *et al.* Transient increase in obese gene expression after food intake or insulin administration. *Nature* **377**, 527–529 (1995).
93. Kiess, W. *et al.* High leptin concentrations in serum of very obese children are further stimulated by dexamethasone. *Horm. Metab. Res.* **28**, 708–710 (1996).
94. Miller, S. G. *et al.* The adipocyte specific transcription factor C/EBP α modulates human ob gene expression. *Proc. Natl Acad. Sci. USA* **93**, 5507–5511 (1996).
95. Devos, P. *et al.* Thiazolidinediones repress ob gene expression in rodents via activation of peroxisome proliferator-activated receptor gamma. *J. Clin. Invest.* **98**, 1004–1009 (1996).
96. Ravussin, E. *et al.* Relatively low plasma leptin concentrations precede weight gain in Pima Indians. *Nature Med.* **3**, 238–240 (1997).
97. Casanueva, F. F. *et al.* Serum immunoreactive leptin concentrations in patients with anorexia nervosa before and after partial weight recovery. *Biochem. Mol. Med.* **60**, 116–120 (1997).
98. Mantzoros, C. *et al.* Cerebrospinal fluid leptin in anorexia nervosa: correlation with nutritional status and potential role in resistance to weight gain. *J. Clin. Endocrinol. Metab.* **82**, 1845–1851 (1997).
99. Maffei, M. *et al.* Absence of mutations in the human ob gene in obese/diabetic subjects. *Diabetes* **45**, 679–682 (1996).
100. Montague, C. T. *et al.* Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* **387**, 903–908 (1997).
101. Strobel, A., Camoin, T. I. L., Ozata, M. & Strosberg, A. D. A leptin missense mutation associated with hypogonadism and morbid obesity. *Nature Genet.* **18**, 213–215 (1998).
102. Clement, K. *et al.* A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* **392**, 398–401 (1998).
103. Clement, K. *et al.* Indication for linkage of the human OB gene region with extreme obesity. *Diabetes* **45**, 687–690 (1996).
104. Reed, D. R. *et al.* Extreme obesity may be linked to markers flanking the human OB gene. *Diabetes* **45**, 691–694 (1996).
105. Comuzzie, A. G. *et al.* A major quantitative trait locus determining serum leptin levels and fat mass is located on human chromosome 2. *Nature Genet.* **5**, 273–276 (1997).
106. Krude, H. *et al.* Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nature Genet.* **19**, 155 (1998).
107. Murphy, J. E. *et al.* Long-term correction of obesity and diabetes in genetically obese mice by a single intramuscular injection of recombinant adeno-associated virus encoding mouse leptin. *Proc. Natl Acad. Sci. USA* **94**, 13921–13926 (1997).
108. Blackburn, G. Effect of degree of weight loss on health benefits. *Obes. Res.* **3** (suppl. 2), 211–216 (1995).
109. Manson, J. E. *et al.* Body weight and mortality among women. *New Engl. J. Med.* **333**, 677–685 (1995).

Acknowledgements. We thank S. Korres for assistance in preparing this manuscript and J. Froude for editing it, and D. Luck, J. Breslow, B. Schneider and M. Stoffel for critically reading it and for thoughtful comments. This work was supported by a grant from NIH/NIDDK (J.M.F.) and in part by an NIH MSTP grant (to J.L.H.).

Correspondence should be addressed to J.M.F. (e-mail: friedj@rockvax.rockefeller.edu).

Hydrological characteristics of the drainage system beneath a surging glacier

Helgi Björnsson

Science Institute, University of Iceland, Dunhaga 3, IS-107 Reykjavik, Iceland

A rare combination of natural circumstances permits assessment of current theories on water flow beneath glaciers. Outburst floods from the subglacial lake Grímsvötn in Iceland took place before, during and after surging of Skeiðarárjökull, the glacier beneath which the outburst floods drain. The observable drainage patterns associated with these floods show the different nature of the basal water conduit system of the glacier during surge and non-surge phases. During surge conditions, basal water is dispersed slowly across the bed in a distributed drainage system; but when the glacier is not surging, water is transported rapidly through a system of tunnels.

Water plays an important role in the dynamics of temperate and subpolar glaciers: by lubricating the glacier base, water facilitates the sliding of ice over its bed. Some water is produced at the bed by frictional and geothermal heat, but the main water supply is melt water from the glacial surface which descends through the glacier to the bed through various internal conduits. Occasionally, water is injected directly at the bed from ice-dammed lakes or during subglacial volcanic eruptions. Drainage along the bed itself is through a system of passageways controlled by the composition and topography of the substratum, the thickness and slope of the overburden ice, and the water supply. Theoretical analyses and field observations suggest the existence of two main types of subglacial drainage systems which may, however, represent the opposite poles of a continuum.

In the first type of system—that which is normally present according to current thinking—water is conducted through tunnels that have been incised upwards into the ice, forming a river-like system of widely spaced tributary branches. These join down-glacier to form a relatively small number of outlet rivers, which emerge from the glacier terminus^{1,2}. In this case, the steady-state water pressure of the passageways varies inversely with water flux¹ because the size of the tunnels adjusts to the flux of water by enlargement of the tunnel walls as a result of frictional melting and contraction caused by inflow of ice. The tunnels transporting large amounts of water grow at the expense of smaller tunnels. However, such ice tunnels cannot adjust to the injection of water over a short time span of hours to days, so there may be a transient increase in water pressure, which reduces basal friction and facilitates sliding. Input of water to this type of drainage system affects sliding of the glacier only within small, local areas.

In the second type of drainage system, water is dispersed across the bed through a distributed network of passageways of variable width. The largest of these may be visualized as water-filled cavities which form behind protuberances in the glacier bed and are hydraulically interconnected, either by small channels cut in the bed, or by narrow ice tunnels located on the downslope side of small steps in the bed. Such chambers lie perpendicular to the ice flow^{3–7}. In this kind of system, water pressure varies directly with the water flux^{5,6}. There is no tendency for larger passageways to collect water and grow at the expense of smaller ones, and water emerges from the glacier margin in many small outlets. Discharge through the system is controlled by the throttling action of the narrow passageways. High pressures are required to drive water through such a drainage system; this lubricates the bed and facilitates sliding over large areas.

We refer to a system of this type as a distributed (or linked-cavity) system.

Both types of drainage system may exist simultaneously beneath glaciers, and may change with seasonal and diurnal inputs of melt water. For instance, a linked-cavity system, conducive to rapid sliding but carrying only small amounts of the melt water, may form temporarily under some parts of a glacier which is otherwise drained by a tunnel system; this formation may be in response to sudden injections of water to which the tunnels cannot immediately adjust.

Glacial surges may occur when the normal, rapid subglacial drainage is disrupted and a linked-cavity system spreads out beneath the glacier and persists for some time^{8–11}. On the basis of observations of long travel times of dye tracers during a surge of the Variegated glacier in Alaska, and of theoretical analyses of a linked-cavity system, Kamb *et al.*⁹ and Kamb⁶ suggested that surging was initiated by a switch from a tunnel system to a distributed system that facilitated sliding. Further, they suggested that the surge continued as long as high water pressure sustained rapid sliding, the sliding itself counteracting the enlargement (by melting) of the orifices connecting the water-filled cavities. Moreover, the observed rapid and concentrated drainage of water from the glacier following the surge was believed to indicate that a tunnel system had been re-established and that this conversion in fact terminated the surge. Kamb's⁶ model predicted that for sufficiently large perturbations, in the form of increased water pressure or decreased sliding, the narrow passageways linking the cavities would grow unstably, and an ice tunnel system would be restored.

Because subglacial drainage systems themselves are inaccessible, and surges infrequent, normal field observations can provide only limited support for models such as Kamb's. However, the massive discharge of water from an ice-dammed lake (a glacier outburst flood or *jökulhlaup*), draining along the glacier bed, provides a unique opportunity to test current hypotheses by comparing outburst floods which occur during glacial surges with those that occur during normal periods. The former are relatively rare; but a series of outburst floods from the subglacial ice-dammed lake Grímsvötn in the interior of the ice cap Vatnajökull (8,100 km²) in Iceland drained beneath the glacier Skeiðarárjökull both before, during and after its 1991 surge (Fig. 1). Observations of these events are consistent with theories of basal drainage through tunnels during the non-surge state and distributed drainage during the surge, as discussed below. They are likewise in accord with the hypothesis that surges are caused by a temporary transition from tunnel to distributed drainage.

Normal drainage of water under Skeiðarárjökull

Melt water from Skeiðarárjökull normally emerges from two main ice tunnels, one on the west side of the terminus and one on the east side, and collects downstream from the glacier margin into the permanent rivers Skeiðará (on the eastern side) and Súla (on the west). Water also collects downstream into the third river channel, Gígja (Fig. 1), from a few small outlets near the centre of the glacier terminus (Fig. 1).

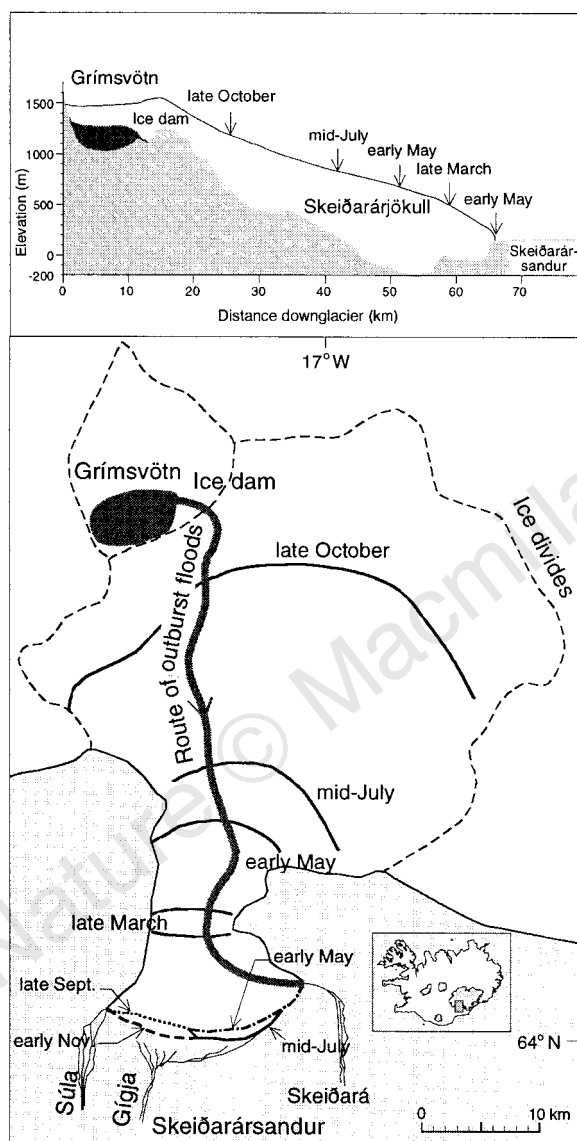


Figure 1 Geographical setting and the progress of the surge of Skeiðarárjökull in 1991. Bottom panel, Location map of the glacier Skeiðarárjökull (1,370 km² in area) which surged in 1991, and the subglacial lake, Grímsvötn, draining outburst floods some 50 km beneath the glacier to river outlets (Skeiðará, Súla and Gígja) on the outwash plain, Skeiðarársandur. A suggested location is shown of an ice tunnel draining glacier outburst floods from Grímsvötn to the river Skeiðará. The gradual enlargement of the surging part of the glacier from the beginning of surging in late March to November 1991, deduced from formation of abundant new crevasses, is shown by lines drawn on the map. The advance of the terminus is shown by lines: the eastern part of the terminus advanced between May and mid-June, and the western part between September and early November. Top panel, a cross-section of Skeiðarárjökull from Grímsvötn down to Skeiðarársandur along the route of the hypothesized ice tunnel. The arrows show propagation of the crevassed area from the start to the end of the surge.

Glacier outburst floods drain regularly from the ice-dammed lake, Grímsvötn, the water flowing beneath Skeiðarárjökull a distance of some 50 km to the terminus at Skeiðarársandur^{12–17}. Such floods from Grímsvötn have occurred at intervals of 1–10 years, with a peak discharge of 600–40,000 m³ s^{−1} at the outwash plain, a duration of 2 days to 4 weeks, and a total volume of 0.5–4.0 km³. All such outburst floods drain through one main ice tunnel which feeds Skeiðará; but if discharge exceeds about 3,000 m³ s^{−1}, water starts to drain from other, smaller tunnels at the central part of the terminus, collecting into the river Gígja. In the most voluminous floods, 10–15 high-capacity ice tunnels develop. The rapid injection of water creates high water pressures, characteristic of distributed systems, as evidenced by water forced up through crevasses on the glacier surface at elevations up to 200 m. A distributed system, however, does not develop, but a tunnel system prevails, and the outburst floods—presumably for this reason—have not led to surges of the glacier. This applies even to the catastrophic flood from Grímsvötn in November 1996, in which 3.2 km³ of water drained from the lake within a period of 40 hours.

Typically, the outlet-river discharge during glacier outburst floods rises approximately exponentially as a function of time, falling rapidly after peaking. Such a discharge pattern can be explained as flow through a single ice tunnel at the glacier bed^{16,18–20}. The flow is controlled mainly by tunnel enlargement due to melting of the walls by frictional heat generated by the flowing water^{1,18}. Occasionally, the temperature of the lake water may be above the melting point of ice, assisting expansion of the tunnel. The start of the drainage from the lake is marked by ice-quakes and subsidence of the lake level; the arrival time of lake water at the outwash plain is defined as the time when a sulphurous odour can be detected from the glacier river, as the lake water contains fluid discharged from a geothermal system situated under the lake.

During the initial phase of an outburst flood from Grímsvötn, water takes about 10 hours to travel the distance (50 km) from the lake down to the glacier margin; but during peak flow, the transit time is estimated at about 3 hours. Average flow velocities thus increase from about 5 m s^{−1} initially to approximately 12 m s^{−1} during the peak flow. To account for the observed peak discharges down the actual gradient of Skeiðarárjökull (0.025), the diameter of the semicircular ice tunnel transporting the water would have to be in the range 25–85 m, assuming a Manning roughness coefficient of 0.08 m^{−1/3} s^{−1} (ref. 16).

Surge-affected drainage from Skeiðarárjökull

In late March 1991, Skeiðarárjökull began to surge. The surge motion was not measured but deduced from the formation of abundant new crevasses, which was monitored from the air. Crevasses initially formed 10 km up-glacier from the terminus and propagated up- and down-glacier (Fig. 1). In early May, a 50-m-high surge front reached the eastern part of the terminus and, after having advanced 450 m, stopped in mid-July. In late September, the western half of the terminus started to advance, and moved forward about 1,000 m before halting in early November. The area affected by increased sliding, producing crevasses, continued to expand up-glacier until the end of October. It ultimately extended some 45 km upwards from the front to within 5 km of the centre of the ice dam containing the Grímsvötn lake. The surge did not produce crevasses in the ice dam or affect its elevation. During the two-stage advance of Skeiðarárjökull, the terminus covered 10 km², and crevasses were produced over an area of about 1,000 km², almost three-quarters of the entire ice drainage basin.

By the start of the melting season in May, it became evident that the outflow-stream pattern of the eastern part of the terminus had changed significantly during the surge. Instead of draining from the normal large outlets, water issued from numerous small streams that developed all along the eastern half of the glacier front. Water was even observed to pour out of crevasses on the glacier surface

tens of metres above the front, indicating a change in the basal drainage system. This water carried high loads of sediment. In contrast, drainage was normal from the western part of the terminus, which had not yet advanced.

In mid-July, when the eastern part of the front stopped advancing, water began again to drain to Skeiðará from one main tunnel, and from a few smaller ice tunnels into the river Gígja. The relative discharge of these two main rivers returned to pre-surge conditions and remained thus throughout the summer.

Drainage of outburst flood water during surge

In mid-September 1991, drainage began from the Grímsvötn lake (Fig. 2). The discharge of the lake water increased approximately exponentially and then fell rapidly; this may be explained as drainage through a single tunnel from the lake^{16,18}. The drainage, however, failed to develop into a normal outburst flood and terminated when less than a quarter of the lake volume had drained (estimated from the subsidence of the lake level, as recorded by precision barometric altimetry, given the known volume distribution by elevation in the lake^{15–17}). Such a truncation of a glacier outburst flood at an early stage has not been observed before.

Water drained out of the lake for 17 days, reached a peak flow rate of $766(\pm 38) \text{ m}^3 \text{ s}^{-1}$ on 29 September (day 13) and then decreased in five days to zero (Fig. 2). The total amount of water drained from Grímsvötn was $3.7(\pm 0.2) \times 10^8 \text{ m}^3$.

On 20 September, four days after this flooding from the lake began, a sulphurous odour from the river Skeiðará indicated that water from Grímsvötn had reached the outwash plain (Fig. 2). Hence, the average flow velocity along the entire route from Grímsvötn was 0.15 m s^{-1} , much lower than the velocity of about 5 m s^{-1} typical of the initial phase of outburst floods from Grímsvötn. The discharge increased to a maximum rate of $320(\pm$

$32) \text{ m}^3 \text{ s}^{-1}$ on 2 October and then declined gradually over three weeks to a minimum ($19 \text{ m}^3 \text{ s}^{-1}$ on 26 October) when $3.3(\pm 0.3) \times 10^8 \text{ m}^3$ of the $3.7(\pm 0.2) \times 10^8 \text{ m}^3$ delivered from Grímsvötn had been released from the surging glacier. No water from Grímsvötn (recognizable by its sulphurous odour) appeared in the other rivers draining the glacier.

The runoff from beneath the glacier to the outwash plain had an extended duration and an attenuated peak, as compared to the outflow from the lake (Fig. 2). The subglacial hydrological system accumulated water continuously until 2 October, when the maximum storage of water at the base was $2.2(\pm 0.26) \times 10^8 \text{ m}^3$.

Swift drainage after the surge of Skeiðarárjökull

Beginning at approximately the end of the surge of Skeiðarárjökull, between late October and late November 1991, a second outburst flood occurred. This was a full event draining Grímsvötn, in which the lake subsided by some 60 m. Discharge in the river Skeiðará showed a continuously increasing flow rate from 27 October, peaked in 26 days (at $2,145(\pm 215) \text{ m}^3 \text{ s}^{-1}$) and receded in 5 days (ref. 21). The discharge of Skeiðará (Fig. 3) was typical for outburst floods from Grímsvötn.

The total volume draining from under the glacier in this full event was $1.3(\pm 0.15) \times 10^9 \text{ m}^3$ in 39 days (from 27 October to 4 December), of which $0.4 \times 10^8 \text{ m}^3$ were a remainder of the first outburst flood, and the average discharge was $455(\pm 45) \text{ m}^3 \text{ s}^{-1}$. The estimated discharge out of Grímsvötn was $1.0(\pm 0.1) \times 10^9 \text{ m}^3$ in 27 days.

Inferences about subglacial drainage systems

Our observations are consistent with theories of distributed drainage during surge, and basal drainage through tunnels during the non-surge state. The response of the surge-state drainage system to the earlier outburst flood offered insights into the nature of the system. Drainage from the lake, Grímsvötn, increased in the manner of a typical glacier outburst flood, indicating normal hydrological conditions at the ice dam. Down-glacier from the dam, by contrast, the observed water transport through Skeiðarárjökull cannot have been in a normal tunnel system.

Along the lines discussed by Kamb *et al.*⁹, we argue that for the observed average water-flow speed of 0.15 m s^{-1} during the initial four days of the outburst flood, a single semicircular ice tunnel with a Manning roughness coefficient $0.08 \text{ m}^{-1/3} \text{ s}^{-1}$ and a hydraulic gradient of 0.025 (the actual gradient of Skeiðarárjökull) would have a diameter of 0.1 m and only carry $\sim 10^{-3} \text{ m}^3 \text{ s}^{-1}$ of water^{16,18}. Hence the observed water discharge must have been transported in a large number of passageways (of the order of 10^4 tunnels, each

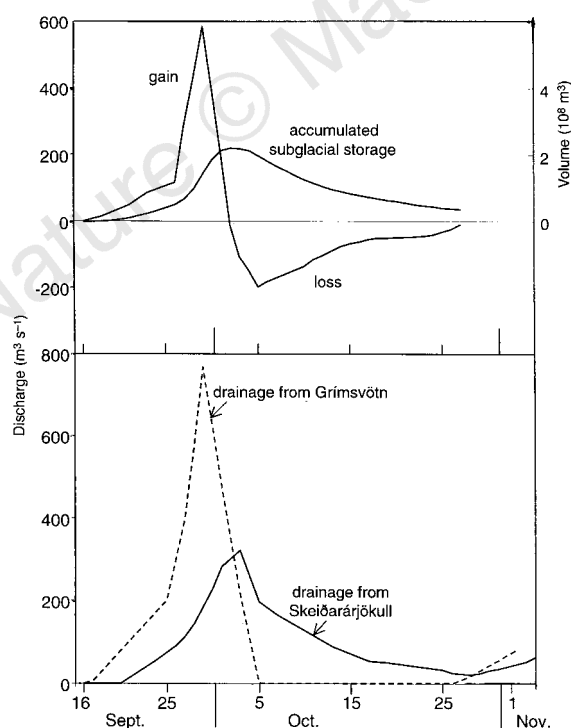


Figure 2 Drainage of outburst flood water during the surge of Skeiðarárjökull. Bottom panel, drainage of outburst flood water from the Grímsvötn lake and from beneath Skeiðarárjökull during its 1991 surge²¹. Comparison of the outflow from the lake and runoff from the glacier terminus. Accuracy is 10%. Top panel, gain, loss and accumulated storage of outburst flood water under or in Skeiðarárjökull during its surge in 1991.

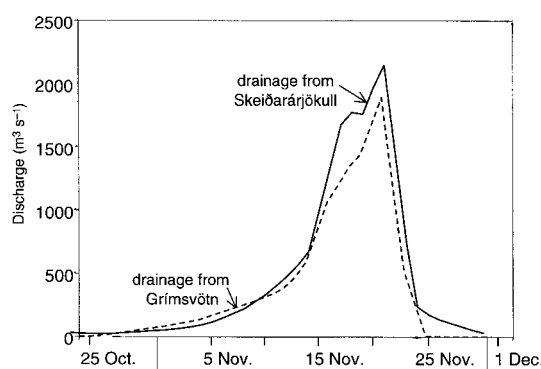


Figure 3 Swift draining of glacier outburst flood water from the Grímsvötn lake and from beneath Skeiðarárjökull after its 1991 surge. Accuracy is 10%. Comparison of the outflow from the lake and runoff from the glacier terminus shows a lag of only a few hours between outflow from the lake and the arrival of the water at the glacier margin.

having a flow velocity of 0.15 m s^{-1} , and a cross-sectional area of 0.006 m^2 , as required to conduct the average discharge of $30 \text{ m}^3 \text{ s}^{-1}$ during the first four days of drainage). Water drained slowly, and spread out into a network of passageways which became larger but which retained the form of a low-transport-cavity distributed system. Water evidently accumulated in storage cavities from which it was then slowly released. The maximum storage of water at the base in early October 1991 was $2.2(\pm 0.26) \times 10^8 \text{ m}^3$ (equivalent to a layer of 30 cm of water evenly distributed beneath the 800 km^2 of the surging area at that time). Cavities in the shape of longitudinally truncated cones, each extending 25 m down-glacier of bumps of height 5 m and occupying 80 km^2 (10% of the bed of the surging glacier) would, for instance, account for the observed maximum storage in early October.

The drainage pattern of the later (November) outburst flood was that of a 'tunnelized' system which rapidly conducted water beneath the glacier from the lake to the margin. There was apparently no significant accumulation of water beneath the glacier (Fig. 3). A flow velocity of about 6 m s^{-1} through a single semicircular ice tunnel of 30 m diameter would account for the peak flow. This indicates that the basal hydrological system up-glacier had switched from a distributed system back to a tunnel system after the surge.

Evidently, the injection of large amounts of water to the base of Skeiðarárjökull, whether from regular spring melting or from outburst floods, normally enhances a tunnel drainage system. During the most voluminous outburst floods, water may break out of the Skeiðará ice tunnel and begin to drain through braided water-courses at high pressures. However, these passageways quickly develop into high-capacity ice tunnels, and outburst floods are not known to have led to surges of Skeiðarárjökull. In contrast, during active sliding, even an outburst flood cannot destroy the distributed drainage system associated with surging. Only when the glacier profile has flattened sufficiently can sliding be reduced so as to permit the re-formation of ice tunnels of sufficient capacity to drain a full-scale outburst flood.

The 1991 surge of Skeiðarárjökull, and the attendant glacier

outburst floods from Grímsvötn late in that year, have produced a variety of phenomena that are consistent with, and which may be explained by, Kamb's model⁶; it would be difficult to explain these events in another way, given our present knowledge. □

Received 1 April; accepted 7 August 1998.

1. Röthlisberger, H. Water pressure in intra- and subglacial channels. *J. Glaciol.* **11**, 177–203 (1972).
2. Shreve, R. L. Movement of water in glaciers. *J. Glaciol.* **11**, 205–214 (1972).
3. Liboutry, L. A. General theory of subglacial cavitation and sliding of temperate glaciers. *J. Glaciol.* **11**, 21–58 (1968).
4. Iken, A. The effect of the subglacial water pressure on the sliding velocity of a glacier in an idealized numerical model. *J. Glaciol.* **27**, 407–421 (1981).
5. Walder, J. S. Hydraulics of subglacial cavities. *J. Glaciol.* **32**, 439–445 (1986).
6. Kamb, B. Glacier surge mechanism based on linked cavity configuration of the basal water conduit system. *J. Geophys. Res.* **92**, 9083–9100 (1987).
7. Walder, J. S. & Fowler, A. Channelized subglacial water channels and cavities. *J. Glaciol.* **40**, 3–15 (1994).
8. Röthlisberger, H. Contribution to discussion of J. Weertman's paper. *Can. J. Earth Sci.* **6**, 941–942 (1969).
9. Kamb, B. *et al.* Glacier surge mechanism: 1982–1983 surge of Variegated Glacier, Alaska. *Science* **227**, 469–479 (1985).
10. Raymond, C. F. & Harrison, W. D. Winter initiation of surges. *Abstr. Hydraulic Effects at the Glacier Bed and Related Phenomena* (ETH Zürich, 1985).
11. Raymond, C. F. How do glaciers surge? A review. *J. Geophys. Res.* **92**, 9121–9134 (1987).
12. Thorarinsson, S. Some new aspects of the Grímsvötn problem. *J. Glaciol.* **14**, 267–274 (1953).
13. Thorarinsson, S. *Vötnin Stríð. Saga Skeiðarárhlaupa og Grímsvatnagosa (The Swift Flowing Rivers. The History of Grímsvötn Jökulhlaups and Eruptions)* (Menningarsjóður, Reykjavík, 1974).
14. Björnsson, H. Explanation of jökulhlaups from Grímsvötn, Vatnajökull, Iceland. *Jökull* **24**, 1–26 (1974).
15. Björnsson, H. *Hydrology of Ice Caps in Volcanic Regions* (Rit 45, Societas Scientiarum Islandica, Reykjavík, 1988).
16. Björnsson, H. Jökulhlaups in Iceland: prediction, characteristics and simulation. *Ann. Glaciol.* **16**, 95–106 (1992).
17. Gudmundsson, M. T., Björnsson, H. & Pálsson, F. Changes in jökulhlaup sizes in Grímsvötn, Vatnajökull, Iceland, 1934–91, deduced from in-situ measurements of subglacial lake volume. *J. Glaciol.* **41**, 263–272 (1995).
18. Nye, J. F. Water flow in glaciers: jökulhlaups, tunnels and veins. *J. Glaciol.* **76**, 181–207 (1976).
19. Spring, U. & Hutter, K. Numerical studies of jökulhlaups. *Cold Reg. Sci. Technol.* **4**, 221–244 (1981).
20. Clarke, G. K. C. Glacier outburst flood from 'Hazard Lake', Yukon Territory, and the problem of flood magnitude prediction. *J. Glaciol.* **28**, 3–21 (1982).
21. Pálsson, S., Zóphóníasson, S., Sigurðsson, O., Kristmannsdóttir, H. & Aðalsteinsson, H. *Skeiðarárhlaup og Framhlaup Skeiðarárjökuls 1991* (Rep. OS-92035/VOD-09 B, Natl Energy Authority, Reykjavík, 1992).

Acknowledgements. The work was supported by The Research Fund of the University of Iceland, the Icelandic Road Authority and the National Power Company of Iceland. I thank B. Kamb for critically reading the manuscript, and my colleague M. Karlsson for assistance with the writing and presentation.

Correspondence and requests for materials should be addressed to the author (e-mail: hb@raunvis.hi.is).

A circumstellar dust disk around a star with a known planetary companion

David E. Trilling & Robert H. Brown

Lunar and Planetary Laboratory, University of Arizona,
1629 East University Boulevard, Tucson, Arizona 85721, USA

A planet with a minimum mass of 0.84 Jupiter masses (M_J) has been indirectly detected¹ in a close orbit (radius 0.11 astronomical units, period 14.65 days) around the star 55 Cancri, which is of spectral type G8 and about 3 billion years old. The detection of excess infrared emission from this system also suggests² the presence of circumstellar dust. Our Solar System has a disk of dust (and larger bodies) that is roughly coplanar with the planets—the so-called Kuiper Belt³. Here we report infrared coronagraphic observations of 55 Cancri in which the light from the primary star is blocked, allowing us to image a circumstellar dust disk. We find that the dust lies in a disk that extends to at least 40 AU, comparable to the expected extent for our Kuiper Belt³, whereas the inferred mass of the disk is approximately ten times that estimated for our Kuiper Belt. The disk around 55 Cancri is relatively dark at a wavelength of 2.3 μm , which is consistent with absorption of light by methane ice on the dust particles. Assuming that the disk is coplanar with the planet, we determine the planet's mass to be $1.9^{+1.1}_{-0.4}$ Jupiter masses. All the available evidence is suggestive of a mature planetary system around 55 Cancri.

Using the CoCo cooled coronagraph with NSFCAM on NASA's Infrared Telescope Facility (IRTF) on Mauna Kea, we searched the 55 Cancri (55 Cnc, ρ^1 Cnc, HD 75732, HR 3522, G8V, 12.53 pc (ref. 4)) system for evidence of circumstellar material. CoCo is a cryogenically cooled infrared Lyot coronagraphic front end to NSFCAM, and uses a gaussian apodized focal plane mask with selectable and articulatable Lyot stops. The coronagraph acts as a two-dimensional Fourier filter, blocking light from the central star as well as light diffracted from the edges of the primary mirror and Cassegrain hole, while allowing imaging to within 18 AU of 55 Cnc. The coronagraphic technique was first used to image circumstellar disks for the Beta Pictoris disk, the first directly imaged circumstellar disk⁵. In these observations, the residual stellar-halo intensity is at most one-tenth of that without the coronagraph. The data reduction requires subtracting the scaled stellar haloes of comparison stars (that are of similar spectral type and magnitude and are nearby on the sky) from the observed stellar halo of 55 Cnc. Halo subtraction was performed on 55 Cnc data by using each of three comparison stars to eliminate the effects of potential anomalies from any one comparison star. This technique (one object star, three comparison stars) was repeated for each of our three filters (H, K' and methane, with central wavelengths of 1.62, 2.12 and 2.28 μm , respectively).

Our final image is shown in Fig. 1. We find a flux excess, not attributable to residual stellar halo, to the northeast and southwest (major-axis direction) of the star, and a smaller flux excess to the northwest and southeast (minor-axis direction) of the star. This signal is most pronounced in the H filter (Fig. 1a), where the signal-to-noise ratio is more than 100; in the K' filter, the signal-to-noise ratio is ~ 50 . Excess circumstellar flux is detected at a very low level ($\sim 3\text{--}5\sigma$) in the methane filter. Low signal in the methane filter is consistent with absorption of light at that wavelength by surfaces composed at least partly of methane ice⁶. Figure 1b shows the H filter data, contoured, with the best-fit ellipse overplotted to accentuate the geometry of the disk. We achieved an H-band sensitivity of 10^{-5} for the ratio of the circumstellar disk flux to the peak intensity of the central star.

We performed a series of tests to ensure that our detected signal was not produced by any systematic error. We have imaged several K-dwarf stars under similar conditions, with the same instrumental configuration, and found no excess flux in any direction beyond the coronagraphic mask. We also imaged a circumstellar disk (MWC480) with known inclination on the sky⁷ and reproduce the known inclination (nearly face-on). We find no evidence of asymmetric flux from the disk around MWC480, thus indicating that any asymmetry detected in the flux from 55 Cnc is endemic to that system only. All this evidence demonstrates that neither the telescope (focus, tracking), the instrumentation, nor the presence of any nearby stars introduced any kind of asymmetric flux excess or bias, and that our data reduction techniques are free of systematic errors that could produce the kind of signal we detected. Comparing the images of 55 Cnc with those of several different standard stars reveals the same asymmetry in all final results; intercomparing two standard stars by the same technique reveals no asymmetries. We therefore believe that the northeast–southwest asymmetric flux excess seen around 55 Cnc is real.

The radial profiles of the disk flux show a power-law dependence with a power law index of -5 to -5.5 (Fig. 2). This index is consistent with light reflected off a population whose spatial extent is similar to that expected for our Kuiper Belt, as solar flux decreases with distance as r^{-2} and the surface density of objects in the Kuiper Belt is estimated to fall off approximately as r^{-3} (ref. 8). Because the signal we detect is light from the central star reflected off dust particles, we can determine a rough relative reflectance spectrum for the circumstellar material (Fig. 3). We find that the material we

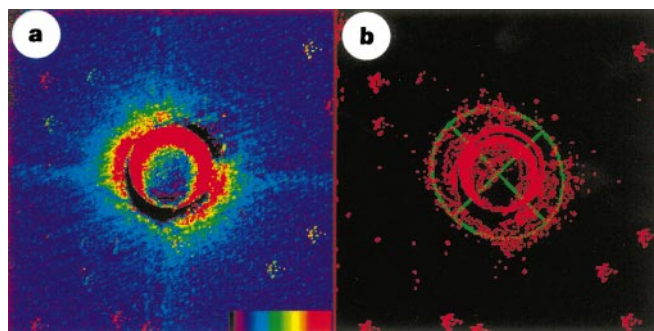


Figure 1 Infrared coronagraphic imaging of circumstellar material around 55 Cnc. **a**, H filter image, 58 min total integration time (29 coadds of 2-min exposures). **b**, H filter total image, contoured, with best-fit ellipse and major and minor axes overlaid. The circle in the centre of the image is the coronagraphic mask, which has a diameter of 3 arcsec. The image is ~ 15 arcsec on a side at a plate scale of ~ 0.06 arcsec per pixel. North is up and east is to the left. The faint vertical and horizontal lines emanating from the coronagraphic mask are the remnants of the diffraction patterns of the telescope spiders, which are used to help to register images. The repeating pattern of spots represents cosmic ray hits in the flat field frame; the clusters result from the slight, repeated shifts used in co-aligning the coronagraphic images, causing a cosmic ray hit to show up as a cluster in the fully processed data. The colour bar in the lower right of **a** has a range of $0\text{--}1.81\text{ mJy arcsec}^{-2}$, with a linear colour map applied in this image. The contour interval in **b** is $0.83\text{ mJy arcsec}^{-2}$, with a lowest contour of $1.11\text{ mJy arcsec}^{-2}$. The ellipse is roughly fitted to the outermost contour drawn in this image. ($1\text{ Jy (Jansky)} = 10^{-12}\text{ W m}^{-2}\text{ }\mu\text{m}^{-1}$ at H band.) The coronagraph blocks light from the central star while allowing imaging to within 1.5 arcsec, or 18 AU for 55 Cnc, of the central star. Interspersed with 55 Cnc observations were observations of comparison stars. All observations were made at less than 2 airmasses, and most observations at less than 1.5 airmasses. Images are sky-subtracted and flat-fielded. A 55 Cnc image was then aligned with a comparison star image, the fluxes were scaled, and the comparison star was subtracted from object star (55 Cnc). In this way, the residual stellar halo was removed from the 55 Cnc images, on the basis of the assumption that the residual stellar halo is simply an expression of the point spread function (PSF), and that the PSF is nearly identical for a nearby comparison star for observations taken at nearly the same time.

detected has a spectrum similar to that of Pluto, often considered the largest and best studied Kuiper Belt object in our Solar System⁹. In particular, the relative reflectance spectra for both the 55 Cnc disk material and Pluto show a drop in reflectance at $\sim 2.32 \mu\text{m}$, where methane ice absorbs photons. Pluto's spectrum returns to a continuum value towards longer wavelengths than the methane ice absorption feature. However, we do not have data for the 55 Cnc circumstellar disk at wavelengths longer than $2.32 \mu\text{m}$, so we can at best say that the spectrum of the disk material is consistent with the presence of a methane ice absorption feature at $2.32 \mu\text{m}$. This spectral signature is consistent with a theoretical extrasolar Kuiper Belt, because methane ice and water ice would be expected to be present at those distances, by cosmochemical arguments¹⁰. Further study of extrasolar Kuiper Belts might prove very useful in discussing circumstellar environments and constraining global Kuiper Belt population distributions and compositions, because the global properties of an extrasolar disk can be studied and mapped in a way in which ours presently cannot.

From the radial profiles of the disk we also measure the apparent detectable extent of the disk. In the major axis direction, we confidently ($>10\sigma$) detect the disk to a distance of at least 40 AU (3.24 arcsec) from the centre of the star, a distance which would be comfortably within our Kuiper Belt³. We measure the extent of the

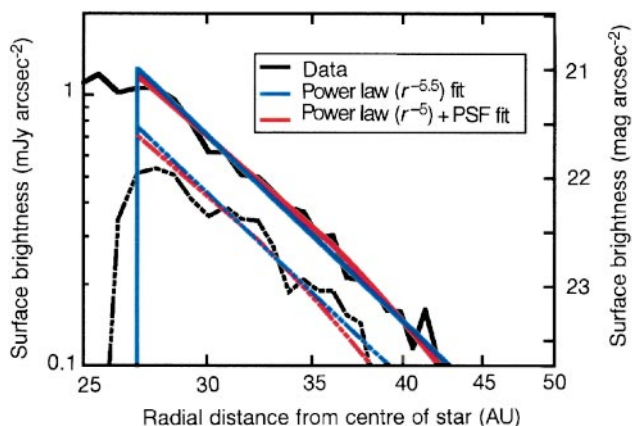


Figure 2 Radial profiles of the disk. These H-band profiles are summed over all phase angles within the major-axis or minor-axis quadrant; the surface brightness of the disk is shown, in mJy (and magnitudes) arcsec^{-2} , in this figure. Data are shown in black, with model disks in blue and red. Solid lines represent the major-axis quadrants, and broken lines signify minor-axis quadrants. The two sets of models drawn represent two slightly different physical scenarios. The blue lines represent a simple power law, with radial dependence $r^{-5.5}$. The red curves represent models in which the disk has an abrupt outer edge at $\sim 35 \text{ AU}$. If the disk were to have this abrupt edge, the point spread function for these observations would smear the outer edge of the disk, making it seem that the disk's extent were larger than its actual size. The model fits for such a scenario are convolutions of a power law (representing the radial structure of and reflected light flux from the disk) and a gaussian with width equal to the seeing (in this case, 0.8 arcsec), representing the observed outward smearing of the edge of the disk. The power law with the best fit in this case has a radial dependence of r^{-5} . In both models, the power-law exponents are in physical agreement with a decrease in solar flux (r^{-2}) and a decrease in population density (r^{-3}) (ref. 8). It is unlikely that a disk would have a truncated outer edge (red curves) without some unusual dynamical cause; it is more likely that the disk's radial structure continues (blue lines) out farther than we can detect it. Nevertheless, the two interpretations fit our data equally well. Note that the same radial functions fit the profiles in both the major-axis and minor-axis quadrants, suggesting that those quadrants represent the same physical environment, with the minor-axis direction foreshortened. Model fits begin at 27 AU , where we are fully confident that the radial extent of the coronagraphic mask is exceeded. The fluctuations from smooth in the data are known artefacts of the polar-coordinate binning that we use; 1σ in our reduced composite frames is $\sim 0.01 \text{ mJy arcsec}^{-2}$ for our H filter data.

semi-minor axis to be $2.88 \pm 0.24 \text{ arcsec}$. From these two measurements, assuming a relatively planar disk³ that is circular and coplanar with the planet's orbit, we find that the inclination of the system on the plane of the sky is $27^{+8}_{-11}^\circ$. Because radial velocity measurements constrain the mass of a planetary companion to be $M_p \sin i = 0.84 M_J$ (ref. 1), we can remove the uncertainty of the planetary companion's mass by measuring the inclination of the planet's orbit on the plane of the sky. We find that the planet's mass is $1.9^{+1.1}_{-0.4} M_J$. The position angle of the disk's major axis is about $50^\circ \pm 10^\circ$.

Our detection of a disk around 55 Cnc joins other measurements (a radial-velocity companion¹, the possibility of a second radial-velocity companion¹, and a $60\text{-}\mu\text{m}$ flux excess²) in indirectly suggesting the presence of a mature planetary system (one that formed from a normal protoplanetary disk and has a number of regular members) around 55 Cnc. By restricting the mass of 55 Cnc b (the planetary companion) to less than $\sim 3 M_J$, we eliminate the possibility that the radial-velocity companion is a brown dwarf or stellar companion, further arguing for a planetary system, not a multiple stellar system. 55 Cnc b has around one-tenth the angular momentum of Jupiter, where most of the angular momentum in our Solar System resides. If, as is assumed, the reflected light we observe from the 55 Cnc disk is primarily reflected from micro-metre-sized particles, the observations require a disk with larger surface area and hence greater mass than that of our Kuiper Belt. Assuming an average particle density of 1 g cm^{-3} , an overall near-infrared reflectance of 6%, as observed for Kuiper Belt objects in our Solar System^{11,12}, and a collisionally evolved population of particles with a power-law index of -3.5 (refs 13, 14), we find that the mass of the disk around 55 Cnc is 10 times the mass of our Kuiper Belt, to within a factor of 3. This circumstellar disk mass is consistent with the estimates of Dominik *et al.*² ($M_{\text{disk}} > 4 \times 10^{-5} M_{\text{Earth}}$) as our Kuiper Belt's mass is estimated to be $4 \times 10^{-2} M_{\text{Earth}}$ (refs 3, 8) and 55 Cnc's circumstellar disk is estimated to be more massive than that by a factor of 10. This inferred mass excess relative to our Solar System's Kuiper Belt (55 Cnc's disk being a factor of 10 more massive) is consistent with the idea that 55 Cnc b migrated inwards from its location of formation¹⁵⁻¹⁷. In this model, a planet migrates

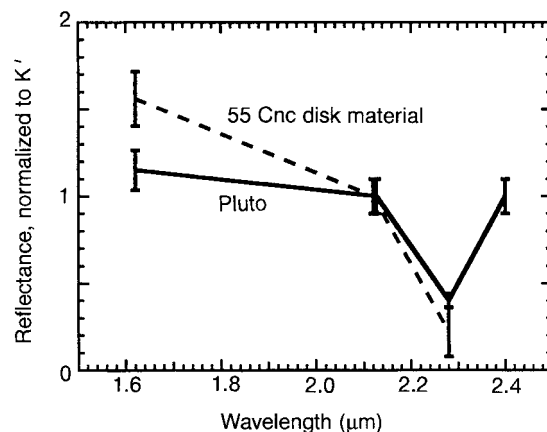


Figure 3 Spectra of 55 Cnc circumstellar disk material and Pluto. This figure shows broadband relative reflectance of the circumstellar material around 55 Cnc and a spectrum of Pluto⁹. Reflectances are normalized to be equal at K' . Relative reflectances for the disk material are found by taking the ratios between images in the three filters. A decrease in reflectance at $2.32 \mu\text{m}$ is clear in both spectra. Pluto's spectrum clearly rises back to a continuum value towards longer wavelengths than $2.32 \mu\text{m}$; our data for 55 Cnc do not extend past the methane filter at $2.32 \mu\text{m}$. Therefore, we cannot confirm the suggested absorption feature, although our data are consistent with the presence of methane ice on the surfaces in the disk around 55 Cnc. Pluto is a relevant comparison both because of heliocentric distance (and hence surface cosmochemistry), and because Pluto is probably genetically and dynamically related to Kuiper Belt objects^{9,18,19}.

inwards by angular momentum exchange with a protoplanetary disk (which initially extends inward nearly to the star). This migration could transfer material from the inner part of the disk to the outer part. The mass enhancement of the outer disk should be related to the angular momentum lost from the planet. Our observations are consistent with a circumstellar disk ten times more massive than our own, consistent with transfer of angular momentum from a migrating planet to the outer disk. Furthermore, the disk of material around 55 Cnc shows an extent consistent with what is expected for our own Kuiper Belt, and a spectral similarity to that expected for our Kuiper Belt. Unlike our Kuiper Belt, a disk around another star can be studied globally to determine its mass, composition and radial structure. Further study of this and other circumstellar disks will allow the characterization of global properties and will in turn lead to an increased understanding of our own Kuiper Belt. □

Received 3 April; accepted 9 July 1998.

- Butler, R. P., Marcy, G. W., Williams, E., Hauser, H. & Shirts, P. Three new '51 Pegasi-type' planets. *Astrophys. J.* **474**, L115–L118 (1997).
- Dominik, C., Laueijs, R. J., Jourdan de Muizon, M. & Habing, H. J. A Vega-like disk associated with the planetary system of ρ^1 Cnc. *Astron. Astrophys.* **329**, L53–L56 (1998).
- Weissman, P. The Kuiper belt. *Annu. Rev. Astron. Astrophys.* **33**, 327–358 (1995).
- The Hipparcos and Tycho Catalogues* (European Space Agency, Noordwijk, The Netherlands, 1997).
- Smith, B. A. & Terrile, R. J. A circumstellar disk around β Pictoris. *Science* **226**, 1421–1424 (1984).
- Brown, R. H., Cruikshank, D. P., Pendleton, Y. & Veeder, G. J. Surface composition of Kuiper belt object 1993SC. *Science* **276**, 937–939 (1997).
- Mannings, V., Koerner, D. W. & Sargent, A. I. A rotating disk of gas and dust around a young counterpart to β Pictoris. *Nature* **388**, 555–557 (1997).
- Duncan, M. J. & Levison, H. F. A disk of scattered icy objects and the origin of Jupiter-family comets. *Science* **276**, 1670–1672 (1997).
- Owen, T. C. et al. Surface ices and the atmospheric composition of Pluto. *Science* **261**, 745–748 (1993).
- Lewis, J. S. *Physics and Chemistry of the Solar System* (Academic, San Diego, 1995).
- Luu, J. X. & Jewitt, D. C. Reflection spectrum of the Kuiper belt object 1993 SC. *Astron. J.* **111**, 499–503 (1996).
- Jewitt, D. C., Luu, J. X. & Chen, J. The Mauna Kea–Cerro Tololo (MKCT) Kuiper belt and centaur survey. *Astron. J.* **112**, 1225–1238 (1996).
- Farinella, P. & Davis, D. R. Short-period comets: primordial bodies or collisional fragments? *Science* **273**, 938–941 (1996).
- Dohnanyi, J. S. Collisional model of asteroids and their debris. *J. Geophys. Res.* **74**, 2531–2554 (1969).
- Lin, D. N. C., Bodenheimer, P. & Richardson, D. C. Orbital migration of the planetary companion of 51 Pegasi to its present location. *Nature* **380**, 606–607 (1996).
- Trilling, D. E. et al. Orbital evolution and migration of giant planets: modeling extrasolar planets. *Astrophys. J.* **500**, 428–439 (1998).
- Murray, N., Hansen, B., Holman, M. & Tremaine, S. Migrating planets. *Science* **279**, 69–72 (1998).
- Duncan, M. J., Levison, H. F. & Budd, S. M. The dynamical structure of the Kuiper belt. *Astron. J.* **110**, 3073–3081 (1995).
- Malhotra, R. The origin of Pluto's peculiar orbit. *Nature* **365**, 819–821 (1993).

Acknowledgements. We thank Doug Toomey for instrumental assistance, and Andy Rivkin for many useful discussions. D.E.T. and R.H.B. were visiting astronomers at the Infrared Telescope Facility which is operated by the University of Hawaii under contract to the National Aeronautics and Space Administration. This work is supported in part by an NSF Graduate Research Fellowship and by various NASA research grants.

Correspondence should be addressed to D.E.T. (e-mail: trilling@lpl.arizona.edu).

Induced magnetic fields as evidence for subsurface oceans in Europa and Callisto

K. K. Khurana*, M. G. Kivelson*†, D. J. Stevenson‡, G. Schubert*†, C. T. Russell*†, R. J. Walker* & C. Polanskey§

* Institute of Geophysics and Planetary Physics, † Department of Earth and Space Sciences, University of California, Los Angeles, California 90095, USA

‡ Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA

§ The Jet Propulsion Laboratory, 4800 Oak Grove Road, Pasadena, California 91109, USA

The Galileo spacecraft has been orbiting Jupiter since 7 December 1995, and encounters one of the four galilean satellites—Io, Europa, Ganymede and Callisto—on each orbit. Initial results from the spacecraft's magnetometer^{1,2} have indicated that neither Europa nor Callisto have an appreciable internal magnetic field, in

contrast to Ganymede³ and possibly Io⁴. Here we report perturbations of the external magnetic fields (associated with Jupiter's inner magnetosphere) in the vicinity of both Europa and Callisto. We interpret these perturbations as arising from induced magnetic fields, generated by the moons in response to the periodically varying plasma environment. Electromagnetic induction requires eddy currents to flow within the moons, and our calculations show that the most probable explanation is that there are layers of significant electrical conductivity just beneath the surfaces of both moons. We argue that these conducting layers may best be explained by the presence of salty liquid-water oceans, for which there is already indirect geological evidence^{5,6} in the case of Europa.

Our insight into the source of the magnetic perturbations

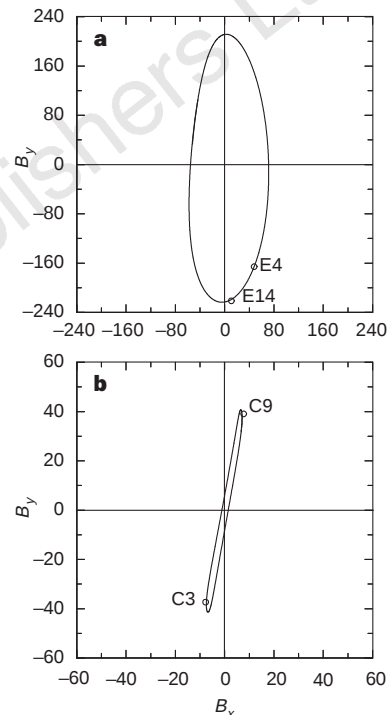


Figure 1 The varying magnetic fields experienced by Europa and Callisto. Near the orbits of the satellites (the orbits of Europa and Callisto lie nearly in Jupiter's spin equator at $9.4 R_J$ and $26.3 R_J$, respectively) but remote from the actual satellite locations, the sources of magnetic field are the internal tilted dipole of Jupiter and the currents flowing in the magnetospheric plasma sheet. The 9.6° tilt between Jupiter's spin and dipole axes implies that the magnetic equatorial plane and the orbital planes of the moons are inclined relative to each other. In a coordinate system with the x axis along the direction of plasma co-rotation, the y axis orientated towards Jupiter, and the z axis along the spin axis of the moon, the z component remains essentially constant. However, the x and y components of the magnetospheric field vary at the synodic period of Jupiter's rotation (11.1 h for Europa and 10.1 h for Callisto) as illustrated in the plots. **a**, The elliptically polarized variation of the magnetic field (in nT) at Europa. Open circles mark the field values corresponding to the E4 and E14 fly-bys. **b**, The almost linearly polarized variation of the magnetic field (in nT) at Callisto. Open circles mark the field values corresponding to the C3 and C9 fly-bys. The time, altitude and latitude relative to the moon's equator for the four passes were: E4, 1996 December 19 06:52:58 UT, 688.1 km, -1.6° ; E14, 1998 March 29 13:21:16 UT, 1,641.3 km, 12.0° ; C3, 1996 November 04 13:24:28 UT, 1,138.9 km, 13.2° ; C9, 1997 June 25 13:47:50 UT, 421.0 km, 2.0° . At the times of these encounters, the SIII west longitude and position relative to the jovian plasma sheet were: E4, 156.8° , $\sim 1 R_J$ above; E14, 184.4° , $\sim 1 R_J$ above; C3, 242.9° , $\sim 1 R_J$ above; C9, 59.9° , $\sim 1 R_J$ below (R_J = radius of Jupiter = 71,492 km). The expected background field was calculated from an empirical model of Jupiter's magnetospheric field that uses spherical harmonics of order 3 to describe the internal field³² and an Euler potential formulation³³ to describe the external field from the current sheet.

recorded near the moons is based on data from four passes for which the signal (induction signature) to noise (perturbations generated within the ambient magnetospheric plasma) ratio is large. Further details of the results from these and other Europa and Callisto fly-bys will appear elsewhere³⁴. Europa and Callisto are located in the inner magnetosphere of Jupiter, where the plasma is confined to a thin sheet (half thickness $\sim 2 R_J$, where R_J is the radius of Jupiter) near the dipole equator. Jupiter's strong magnetic field keeps the ambient plasma close to rigid co-rotation with Jupiter, implying that it overtakes the orbiting moons from behind. In the rest frame of the moons, the magnetospheric field wobbles as shown in Fig. 1. A varying magnetic field with a peak amplitude of ~ 220 nT (~ 40 nT) is imposed on Europa (Callisto) at the synodic period (the period observed in the moon's rest frame) of Jupiter's rotation. Conductors within or surrounding the moons respond to such varying fields by generating eddy currents on their surfaces. In a uniform oscillating field, eddy currents on the surface of a highly conducting sphere or a spherical shell generate the field of an oscillating magnetic dipole external to the conductor and cancel the oscillating field inside the conductor⁷. Continuity of the normal component of the magnetic field requires that at the pole of the induced dipole, the induced field cancel the background field.

The interactions of the moons with the magnetospheric plasma perturb the background field thereby complicating the interpretation of the induction signature. At Europa, the principal plasma

effect comes from the mass loading of the plasma from newly picked-up water group ions. Such an interaction enhances the field strength upstream of the moon and decreases it downstream of the moon. Other plasma-related effects that obscure the induction signature include the standing Alfvén wave current system that flows through an external conducting layer⁸ surrounding the moon (an ionosphere, for example), diamagnetism from newly picked-up plasma, an expansion fan introduced in the wake of a non-conducting moon by the absorption of plasma by the moon⁹, and the ambient ultra-low frequency (ULF) waves present in the plasma sheet of Jupiter¹⁰. To minimize the complications arising from the plasma effects, we have concentrated on two of the Europa fly-bys (orbits E4 and E14) and two of the Callisto fly-bys (orbits C3 and C9) for which the moons were located outside the central dense part of Jupiter's plasma sheet. For these passes, the moon-plasma interaction was weak and the background fluctuations from ULF waves were small.

Figure 2 shows data from encounter E14. (See Fig. 1 and Fig. 1 legend for parameters of this and other encounters.) Also plotted are the predictions from the induction model (with no adjustable parameters). Near the equatorial plane of Europa, where these observations were made, induction is not expected to modify greatly the component of the magnetic field in the z -direction, B_z . We believe that the field magnitude and B_z are enhanced during this and other Europa encounters principally by mass loading¹¹, diversion

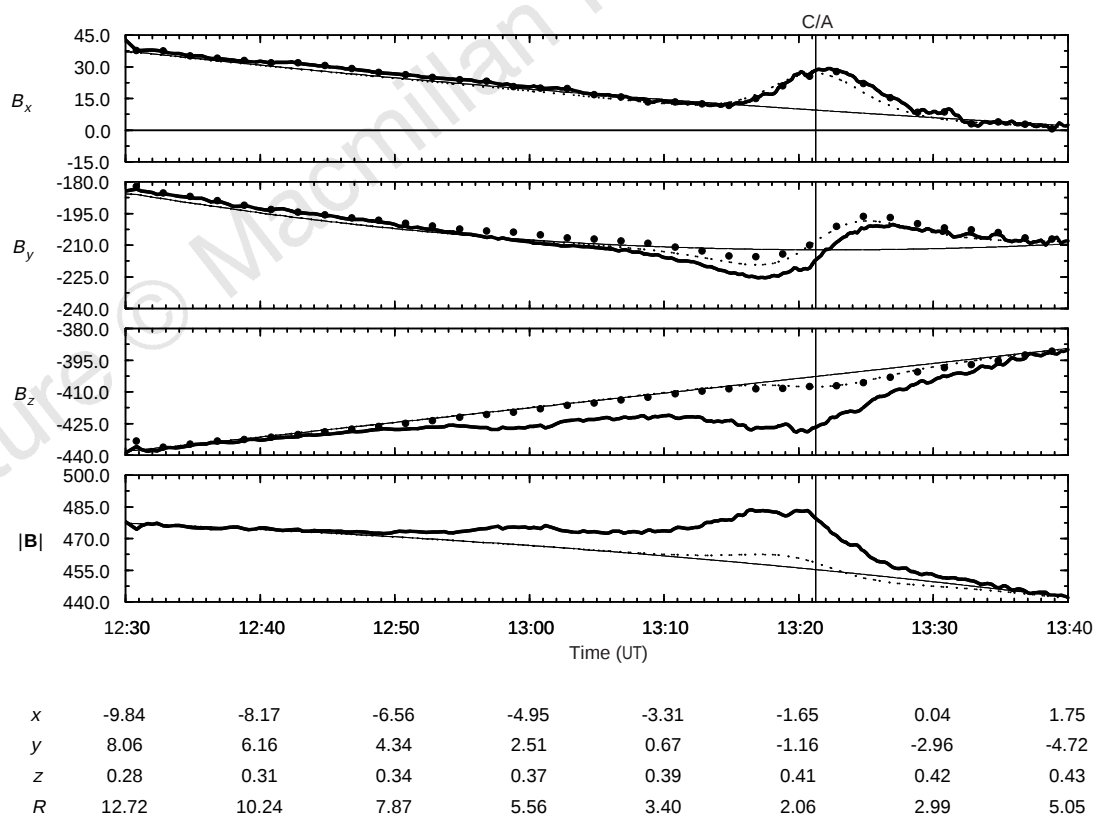


Figure 2 Magnetic field observations for the E14 pass. The plot covers a time interval of 70 min during which the spacecraft moved inward from an initial range of $13 R_E$ to $\sim 2 R_E$ at closest approach (C/A), and travelled back out to a distance of $5 R_E$ from Europa (here R_E is the radius of Europa). The observed magnetic field components and magnitude (in nT) are plotted as thick solid lines. The thin solid lines represent the estimated background field of Jupiter's magnetosphere along Galileo's trajectory, estimated from the interpolation of magnetic data obtained along the trajectory when the spacecraft was sufficiently far from Europa ($> 12 R_E$) that the induction and plasma interaction effects were negligible. The modelled field (induction + background) is shown by dotted lines, and provides a satisfactory fit to the large-scale field variations. Filled circles

show the observations corrected for the plasma pick-up effect described in the text, and this correction is seen to improve the agreement. The variables x , y and z shown under the figure describe the trajectory of the spacecraft in the coordinate system described in Fig. 1 legend. R is the range of the spacecraft. All distances are in units of R_E ($1 R_E = \text{radius of Europa} = 1,560$ km). The data sampling rate changes from 25 s to 1 s at 13:05:40 UT. The induced model field used here and in Fig. 3 was calculated using the equations $B_r = B_0(t)(1 - (a/r)^3)\cos\theta$, $B_\theta = B_0(t)(1 + 0.5(a/r)^3)\sin\theta$, $B_\phi = 0$, where B_r , B_θ and B_ϕ are the components of magnetic field in a spherical coordinate system which has a pole direction antiparallel to the background field, B_0 .

of flow by the conducting obstacle and associated plasma effects. As the component of the magnetic field (B_x) along the flow direction is small ($B_x/|B| \ll 1$) for all the encounters, plasma effects will be symmetric about a plane through the centre of the moon perpendicular to the background magnetic field. Above and below this plane, plasma currents drape the field around the moon, causing bending. In the symmetry plane, near which these observations were made, plasma effects change the field strength without changing its orientation. If the orientation of the field is not to change, each component must change by the same fractional amount: $\delta B_x/B_x \approx \delta B_y/B_y \approx \delta B_z/B_z \approx \delta|B|/|B|$. Here $\delta|B|$ is the change in the field strength. Thus, by reducing each component by a factor $(1 - \delta|B|/|B|)$, we can approximately remove the plasma contributions. The correction improves the agreement between the observations and the model for both the E14 and E4 (not shown here) fly-bys.

At Callisto, corrections for plasma effects are not needed. Figure 3 shows the observed perturbations and the induced dipole model for the C3 and C9 passes: agreement is good. As these Callisto encounters occurred at opposite phases of the variation of the background field (away from Jupiter for the C3 fly-by and towards Jupiter for C9), the induced dipole moments were roughly antiparallel (see Fig. 3a and b). This convincingly demonstrates that the Callisto observations cannot be explained by a fixed internal dipole. For the multiple Europa observations, the orientation of the time-varying component of Jupiter's field changed only slightly among the relevant passes, so the induced dipole moments differed only slightly. A fixed internal dipole cannot be excluded, although its orientation at a large angle to the spin axis seems improbable.

The induced-field model for Europa and Callisto constrains their interior structures by requiring conducting paths at or near the surfaces. It is known that a periodically varying magnetic field (angular frequency ω) acting on an electrically conducting object of conductivity σ decays in an e-folding length of $S = (\omega\mu\sigma/2)^{-1/2}$; here S is the skin depth and μ is the permeability. If the period of the

wave is 11 h and the conductivity is 1 S m^{-1} , $S \approx 95 \text{ km}$. The solution for a spherical shell can be expressed in terms of Bessel functions, but when the thickness of the conducting layer is $\geq 0.1 S$ and $S \ll a$ with a the radius of the conductor, the solution outside the conductor is the sum of an induced dipole field (whose surface field at the pole is equal and opposite to the background field) and the uniform background field⁷ (see Fig. 2 legend).

The observed amplitudes of the induced signatures of Europa and Callisto require conducting layers of depth $>0.1 S$ near the surfaces. For Europa, an obvious candidate for conducting paths is its ionosphere¹² or a cloud of pickup ions¹³. However, estimates of the conductivity above the surface give skin depths for a ten-hour wave much larger than the moon itself. Thus the wave easily penetrates the ionosphere without causing significant induction. Skin depths (for an approximately ten-hour wave) of various materials likely to be found in the icy outer layers of Europa or Callisto can be determined. A rocky mantle composed of (pure) ice and rocks would have a skin depth greater than 10^6 km . Metals such as iron are not expected to be abundant in the outer layers of a differentiated body. Induction from inner metallic cores can also be ruled out. A metallic core whose radius is half the moon's radius would produce a signature that is only one-eighth as large as observed because the induced dipole field magnitude falls as inverse distance cubed. An ocean whose salinity is comparable with Earth's ocean could produce the signature. The conductivity of Earth's ocean water¹⁴ (salinity 3.75%) is $\sim 2.75 \text{ S m}^{-1}$ at 0°C . Thus Earth-like oceans with thicknesses $>10 \text{ km}$ could generate the observed signatures in Europa and Callisto. The conductivity of ocean water is electrolytic and requires only small amounts (a few per cent) of dissolved salts (like NaCl) or acids (like H_2SO_4) that hydrolyse readily.

Induced fields at Europa have been considered since 1985¹⁵, followed by more recent speculations¹⁶. Neubauer¹³ noted that the published^{1,2} dipole moments of the magnetic field perturbations near Europa and Callisto could be fully or partially explained by induction from subsurface oceans or dirty ice layers near the melting point.

The possibility of a liquid-water ocean beneath the icy surface of Europa has been debated for more than two decades. Accretional and radiogenic heat sources are large enough to dehydrate the interior of Europa early in its evolution, leaving the moon covered with a layer of liquid water $\geq 100 \text{ km}$ thick¹⁷. Measurements by the spacecraft Galileo of Europa's gravitational field show that Europa is strongly differentiated (with a metallic core), and that it indeed has a water ice-liquid outer layer $\sim 100 \text{ km}$ thick¹⁸.

Early thermal models considered only the conductive cooling and freezing with time of the outer layer of water, and predicted that at present liquid water existed beneath an ice shell. It was later shown¹⁹ that with thickening, the outer layer of ice would become unstable to thermal convection, promoting heat transfer through the ice and solidification of the underlying water. Complete freezing of the outer layer of water in a small fraction of geological time is possible but not certain²⁰, even for pure water. Additionally, tidal dissipation in Europa's ice shell provides a heat source that could offset the convective cooling of the ice and prevent complete solidification of the water ocean²¹.

The competition between the tendency of tidal heating to maintain a liquid water ocean and that of ice convection to freeze the ocean has been analysed, but a definitive conclusion was not reached²²⁻²⁴. Significant uncertainties include the unknown rheology of ice²⁵ and the dependence of the thermal conductivity of ice on its temperature and physical state. A thermally insulating surface layer would promote stabilization of a liquid-water ocean²³. The occurrence of minor constituents in the ice and ocean, such as salts²⁶ and ammonia, would affect the rheology of the ice and the freezing temperature of the ocean. Tidal heating of the main faults in the ice²⁷ and frictional dissipation due to forced circulation in a thin liquid-water ocean may be important.

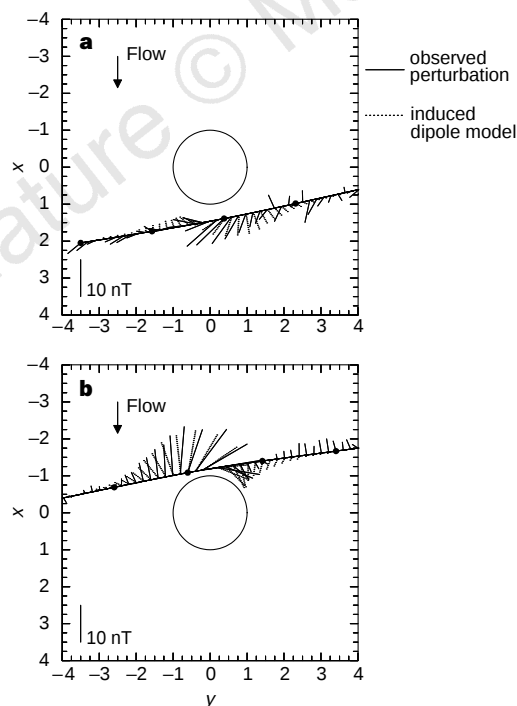


Figure 3 Magnetic field observations from the C3 and C9 passes. **a**, The magnetic field perturbations (vectors drawn with solid lines) and the modelled induction field (vectors shown dotted) along the trajectory of the C3 encounter in the x-y plane. **b**, The magnetic field perturbations and the modelled induction field for the C9 encounter. The distance scale is in units of R_C ($1R_C$ = radius of Callisto = 2,409 km).

Whereas the possible existence of a liquid-water ocean on Europa is plausible, the opposite is true for Callisto. Callisto consists of roughly equal amounts of rock and ice. It is not tidally heated, there is no geological evidence for significant endogenic modification of its surface, and gravity measurements by the Galileo spacecraft show only partial separation of the ice and rock in its interior²⁸. These observations are consistent with little modification of Callisto since its accretion. We note that thermal models of Callisto give no hint of a subsurface liquid-water ocean¹⁷. Accretional and radiogenic heating are marginally able to separate the ice and rock inside Callisto, but the present gravitational evidence shows that unlike Ganymede²⁹ separation has been incomplete: Callisto has not been heated enough to have melted all its ice.

Is it possible for some of the ice in the outer part of Callisto to have melted, and could a near-surface liquid-water layer be prevented from freezing? Because accretional heating is largest when a planet is near maximum size, we consider that it is possible for the ice to have melted in the outer layers of Callisto. More problematic is keeping such a layer from freezing; tidal heating is necessary for the maintenance of a liquid ocean on Europa, and there is no tidal heating on Callisto. The presence of 'antifreeze' (salts or ammonia) would help. The layer needs to have substantial thickness, and for this reason an ocean separating two solid convecting regions is most plausible. The possibility of a liquid-water ocean in Callisto is startling, but we have no other explanation for the near-surface highly electrically conducting layer required by the observed induction signal. Of the two icy galilean satellites, it would be more plausible for Ganymede to have a subsurface liquid-water ocean. Ganymede is completely differentiated; extensive endogenic modification of its surface and the existence of an intrinsic magnetic field³ imply a dynamic interior in the past, and possibly also in the present³⁰. Perhaps Ganymede also has an internal liquid-water ocean if Callisto has one, but Ganymede's intrinsic magnetic field obscures any induction signal.

One possible source of heat not yet considered is ohmic heating by the eddy currents in the moons. The dissipated power can be estimated from the expression³¹ $\text{power/area} = \omega B^2 / 4\mu_0$, which is the ohmic loss from a propagating electromagnetic wave in a conducting waveguide (here μ_0 is the permeability of vacuum). Multiplication by the surface area of the moon and substitution of varying field amplitudes of 220 nT (Europa) and 40 nT (Callisto) gives 5×10^6 (S/100 km) W for Europa and 4×10^5 (S/100 km) W for Callisto; nominal values for S are of the order of 100 km. More rigorous estimates would not change the conclusion that the heat input from this source is negligible.

We conclude from an analysis of the magnetic field observations that it is very likely that both Europa and Callisto possess internal salty liquid-water oceans. In the case of Europa, this conclusion is supported by indirect geological evidence^{5,6}. □

Received 24 June; accepted 14 August 1998.

- Kivelson, M. G. *et al.* Europa's magnetic signature: report from Galileo's first pass on December 19, 1996. *Science* **276**, 1239–1241 (1997).
- Khurana, K. K., Kivelson, M. G., Russell, C. T., Walker, R. J. & Southwood, D. J. Absence of an internal magnetic field at Callisto. *Nature* **387**, 262–264 (1997).
- Kivelson, M. G. Discovery of Ganymede's magnetic field by the Galileo spacecraft. *Nature* **384**, 537–541 (1996).
- Kivelson, M. G. *et al.* A magnetic signature at Io: initial report from the Galileo magnetometer. *Science* **273**, 337–340 (1996).
- Carr, M. H. *et al.* Evidence for a subsurface ocean on Europa. *Nature* **391**, 363–365 (1998).
- Pappalardo, R. T. *et al.* Geological evidence for solid-state convection in Europa's ice shell. *Nature* **391**, 365–368 (1998).
- Parkinson, W. D. *Introduction to Geomagnetism* 308–340 (Scottish Academic, Edinburgh, 1993).
- Neubauer, F. M. Nonlinear standing Alfvén wave current system at Io: Theory. *J. Geophys. Res.* **85**, 1171–1178 (1980).
- Owen, C. J. *et al.* The lunar wake at 6.8 R_L : WIND magnetic field observations. *Geophys. Res. Lett.* **23**, 1263–1266 (1996).
- Khurana, K. K. & Kivelson, M. G. Ultra low frequency MHD waves in Jupiter's middle magnetosphere. *J. Geophys. Res.* **94**, 5241–5254 (1989).
- Ip, W.-H. Europa's oxygen exosphere and its magnetospheric interaction. *Icarus* **120**, 317–325 (1996).
- Kliore, A. J., Hinson, D. P., Flasar, F. M., Nagy, A. F. & Cravens, T. E. The ionosphere of Europa from Galileo radio occultations. *Science* **277**, 355–358 (1997).
- Neubauer, F. M. The subalvénic interaction of the Galilean satellites with the Jovian magnetosphere. *J. Geophys. Res. (Planets)* **103**, 19843–19866 (1998).

- Montgomery, R. B. in *American Institute of Physics Handbook* 125–127 (McGraw Hill, New York, 1963).
- Colburn, D. S. & Reynolds, R. T. Electrolytic currents in Europa. *Icarus* **63**, 39–44 (1985).
- Kargel, J. S. & Consolmagno, G. J. Magnetic fields and the detectability of brine oceans in Jupiter's icy satellites. *Lunar Planet. Sci. Conf. XXVII*, 643–644 (1996).
- Schubert, G., Spohn, T. & Reynolds, R. in *Satellites* (eds Burns, J. A. & Matthews, M. S.) 224–292 (Univ. Arizona Press, Tucson, 1986).
- Anderson, J. D. Europa's differentiated internal structure: inferences from four Galileo encounters. *Science* (in the press).
- Reynolds, R. T. & Cassen, P. On the internal structure of the major satellites of the outer planets. *Geophys. Res. Lett.* **6**, 121–124 (1979).
- Kirk, R. L. & Stevenson, D. J. Thermal evolution of a differentiated Ganymede and implications for surface features. *Icarus* **69**, 91–135 (1987).
- Cassen, P. M., Reynolds, R. T. & Peale, S. J. Is there liquid water on Europa? *Geophys. Res. Lett.* **6**, 731–734 (1979).
- Squyres, S. W., Reynolds, R. T., Cassen, P. M. & Peale, S. J. Liquid water and active resurfacing on Europa. *Nature* **301**, 225–226 (1983).
- Ross, M. N. & Schubert, G. Tidal heating in an internal ocean model of Europa. *Nature* **325**, 133–134 (1987).
- Ojakangas, G. W. & Stevenson, D. J. Thermal state of an ice shell on Europa. *Icarus* **81**, 220–241 (1989).
- Durham, W. B., Kirby, S. H. & Stern, L. A. Creep of water ices at planetary conditions: a compilation. *J. Geophys. Res.* **102**, 16293–16302 (1997).
- McCord, T. B. *et al.* Salts on Europa's surface detected by Galileo's Near Infrared Mapping Spectrometer. *Science* **280**, 1242–1245 (1998).
- Stevenson, D. J. in *Proc. Europa Conf.* 69–70 (San Juan Capistrano Research Inst., San Juan Capistrano, CA, 1996).
- Anderson, J. D. *et al.* Distribution of rock, metals, and ices in Callisto. *Science* **280**, 1573–1576 (1998).
- Anderson, J. D., Lau, E. L., Sjogren, W. L., Schubert, G. & Moore, W. B. Gravitational constraints on the internal structure of Ganymede. *Nature* **384**, 541–543 (1996).
- Schubert, G., Zhang, K., Kivelson, M. G. & Anderson, J. D. The magnetic field and internal structure of Ganymede. *Nature* **384**, 544–545 (1996).
- Jackson, J. D. *Classical Electrodynamics* 2nd edn 338 (Wiley, New York, 1975).
- Connerney, J. E. P. Magnetic fields of the outer planets. *J. Geophys. Res.* **98**, 18659–18679 (1993).
- Khurana, K. K. Euler potential models of Jupiter's magnetospheric field. *J. Geophys. Res.* **102**, 11295–11306 (1997).
- M. G. Kivelson *et al.* Europa and Callisto; induced or intrinsic fields in a periodically varying plasma environment. *J. Geophys. Res.* (submitted).

Acknowledgements. We thank D. Bindschadler of the Jet Propulsion Laboratory for support in all phases of data collection and J. Mafi for data processing and preparation of data plots. This work was partially supported by the Jet Propulsion Laboratory.

Correspondence and requests for materials should be addressed to K.K.K. (e-mail: kkhurana@igpp.ucla.edu).

Quantized conductance through individual rows of suspended gold atoms

Hideaki Ohnishi*, Yukihito Kondo* & Kunio Takayanagi*†

* Takayanagi Particle Surface Project, ERATO, Japan Science and Technology Corporation, 3-1-2 Musashino, Akishima, Tokyo 196, Japan

† Department of Material Science and Engineering, Tokyo Institute of Technology, 4259 Nagatsuta, Midori-ku, Yokohama, Kanagawa 226, Japan

As the scale of microelectronic engineering continues to shrink, interest has focused on the nature of electron transport through essentially one-dimensional nanometre-scale channels such as quantum wires¹ and carbon nanotubes^{2,3}. Quantum point contacts (QPCs) are structures (generally metallic) in which a 'neck' of atoms just a few atomic diameters wide (that is, comparable to the conduction electrons' Fermi wavelength) bridges two electrical contacts. They can be prepared by contacting a metal surface with a scanning tunnelling microscope (STM)^{4–7} and by other methods^{8–12}, and typically display a conductance quantized in steps of $2e^2/h$ ($\sim 13 \text{ k}\Omega^{-1}$)^{13,14}, where e is the electron charge and h is Planck's constant. Here we report conductance measurements on metal QPCs prepared with an STM that we can simultaneously image using an ultrahigh-vacuum electron microscope, which allows direct observation of the relation between electron transport and structure. We observe strands of gold atoms that are about one nanometre long and one single chain of gold atoms suspended between the electrodes. We can thus verify that the conductance of a single strand of atoms is $2e^2/h$ and that the conductance of a double strand is twice as large, showing that

equipartition holds for electron transport in these quantum systems.

Our miniaturized STM was placed at the specimen position of an ultrahigh-vacuum (UHV), high-resolution transmission electron microscope (TEM) which was equipped with a field emission gun operated at 200 kV. Operating this UHV TEM at 10^{-8} Pa allowed the clean fabrication of gold tips and samples *in situ*, thereby eliminating any effects of contamination on the structures and conductances.

A mechanically sharpened gold tip was positioned close to a gold island that had been deposited on a thin copper wire (the electrical connections are shown in Fig. 1). This gold tip could be displaced in three dimensions using a shear-type piezoelectric transducer. The gold tip was dipped into the gold island and slowly withdrawn at constant speed under computer control. Structural changes were observed continuously during withdrawal by means of a video monitor, and video images were recorded at 33-ms intervals; conductance was measured simultaneously. Stepped changes in conductance, corresponding to changes in the thickness of the gold bridge, were observed during withdrawal that were similar to those reported for STM experiments^{4–7}, a mechanically controllable break technique⁸, pin-plate experiments⁹ and switching electrodes^{10–12}.

Figure 2a–f shows electron microscope images of a gold bridge formed between the substrate and tip during withdrawal of the tip. The tip and substrate are oriented in the [110] direction and 0.2-nm spacings of the (200) lattice planes can be seen. The dark lines in the gold bridges represent rows of gold atoms spanning the distance from the substrate to the tip. The bridge thins in Fig. 2a–e and has broken in Fig. 2f; it appears to have reorganized itself into stable structures during withdrawal, with lines disappearing one-by-one, so that each bridge maintains a regular and uniform thickness. The structure of these thin gold bridges and their relation to conductance is a subject for future study. At this point, it should simply be noted that the lattice in Fig. 2a has a dislocation and that the gap between the dark lines of Fig. 2d is greatly enlarged relative to the (200) lattice spacings (0.2 nm) of the tip and substrate. The enlarged gap between the two rows is not an imaging artefact. We confirmed from the image simulation that TEM images reproduce the same

gap distance as the model structure with two separate atomic rows (lattice planes).

More than 300 experiments provided clear evidence that linear strands of gold atoms make up the bridges and that the strands have a unit conductance of $2e^2/h$. Figure 3a shows the change in conductance during the withdrawal of an STM tip. The conductance was sampled at 33-ms intervals in order to coincide with the video frames. The conductance changed stepwise, with a step height corresponding to the unit conductance $G_0 = 2e^2/h$. TEM images corresponding to steps A and B in Fig. 3a are shown in the left and right panels of Fig. 3b, respectively. Both images show only a single dark line spanning the gap between the tip and the substrate. The fact that the conductance of step A was twice that of step B was explained by analysing the intensity profiles perpendicular to the bridges. As shown by the intensity profiles in Fig. 3c, the profile peak (hatched area) of the left bridge is twice the height of the right bridge. According to electron microscopy imaging theory¹⁵, the peak height corresponds to the number of atoms stacked in the depth direction. We conclude that the right bridge in Fig. 3b is a single row of atoms having a unit conductance of $2e^2/h$ and that the left bridge contains two rows with a conductance equal to twice the unit conductance (corresponding models are shown in Fig. 3d).

We next investigated the structure of a bridge of atoms and of individual gold atoms arrayed in a linear strand (the electron microscope images the positions of the nuclei of the atoms). A very thin gold film was set in the specimen stage of the UHV electron microscope and bombarded with a very strong ($\sim 100 \text{ A cm}^{-2}$) electron beam in order to bore holes in the film. A free-standing nanobridge of gold was formed in vacuum by further thinning the material between adjacent holes. A nanobridge of 10 nm long and with four atomic rows (or planes) in its diameter was formed in the [110] direction of the (001) film¹⁶. From the (110) films, a single strand in the [001] direction frequently formed. This strand of atoms was often observed continuously for two minutes or more before breaking (a photograph of one such strand is shown in Fig. 4). Between the coloured edges of the gold film, four coloured dots corresponding to four gold atoms of the strand can be recognized. The gold atoms are spaced at 0.35–0.40 nm intervals which are much larger than the nearest-neighbour spacing (0.288 nm) of gold

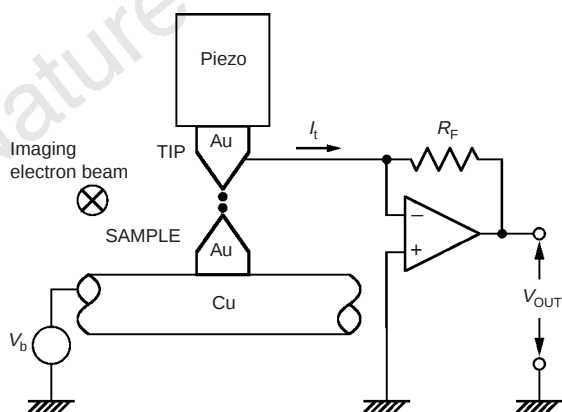


Figure 1 Scanning tunnelling microscope (STM) configuration built at the specimen stage of a UHV electron microscope. The conductance, $G = I_t/V_b$, was obtained by measuring $V_{out} = -R_F I_t$ for a constant bias voltage V_b , where R_F is the feedback resistor for current sensing, and I_t is the current passing through the contact between the tip and the sample. For most experiments, $V_b = 13 \text{ mV}$ and $R_F = 100 \text{ k}\Omega$. The imaging electron beam entered from top and typically had a current density of 0.45 fA nm^{-2} at the contact between the tip and the substrate. Bridges of gold atoms formed at the contact were imaged at $\times 10^7$ magnification on a video monitor and simultaneously recorded with the current at intervals of 33 ms.

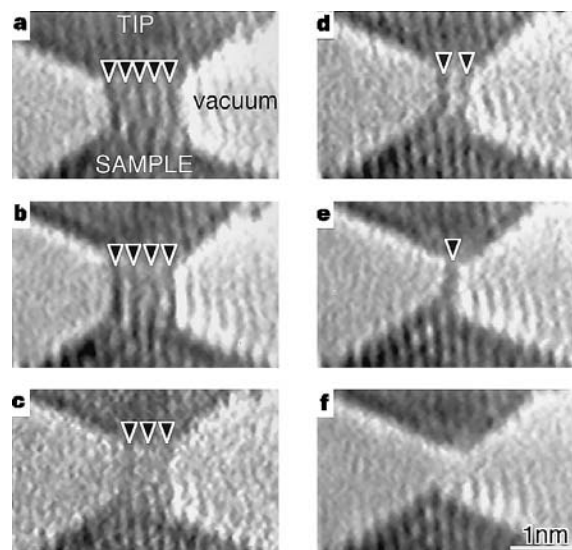


Figure 2 Electron microscope images of a contact while withdrawing the tip. A gold bridge formed between the gold tip (top) and gold substrate (bottom), thinned from a to e and ruptured at f, with observation times of 0, 0.47, 1.23, 1.33, 1.80 and 2.17 s, respectively. Dark lines indicated by arrowheads are rows of gold atoms. The faint fringe outside each bridge and remaining in f is a ghost due to interference of the imaging electrons. The conductance of the contact is 0 at f and $\sim 2 \times (13 \text{ k}\Omega)^{-1}$ at e. $V_b = -10 \text{ mV}$ and $R_F = 10 \text{ k}\Omega$.

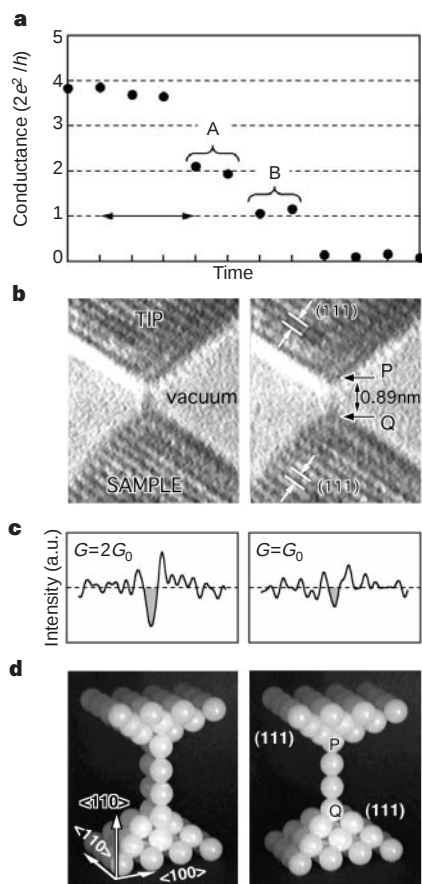


Figure 3 Quantized conductance of a single and a double strand of gold atoms. **a**, Conductance change of a contact while withdrawing the tip. Conductance is shown in units of $G_0 = 2e^2/h \sim (13 \text{ k}\Omega)^{-1}$. $V_b = +13 \text{ mV}$ and $R_F = 100 \text{ k}\Omega$. **b**, Electron microscope images of gold bridges obtained simultaneously with the conductance measurements in **a**. Left, bridge at step A; right, bridge at step B. **c**, Intensity profiles of the left and right bridges shown in **b**. The shaded area is the intensity from the bridge after subtraction of the background noise. **d**, Models of the left and right bridges. The bridge at step A has two rows of atoms; the bridge at step B has only one row of atoms. The distance from P to Q (see **b**) is about 0.89 nm , wide enough to have two gold atoms in a bridge if the gold atoms have the nearest-neighbour spacing of the bulk crystal (0.288 nm).

atoms in the crystal.

The study of electron transport through point contacts in metals dates back to Maxwell's studies of the diffusive transport regime¹⁷ and to Sharvin's proposal for the ballistic transport regime¹⁸. Quantization of conductance in the quantum mechanical ballistic regime was demonstrated experimentally for point contacts of two-dimensional GaAs–AlGaAs heterostructures^{13,14} and for metal QPCs in the STM^{4–7} and other geometries^{8–12}, and was explained by Landauer's formula¹⁹ and Kubo's work²⁰. Our result proves quantization of conductance for a linear strand of gold atoms. When a linear strand connects two reservoirs biased by $dE (= eV_b)$ to produce a current $dI (= I_c)$, then $dI/dE = 2e/h$. There are two foundations for this. First, electrons that move in a linear chain of gold atoms, with a momentum in the range $(p, p + dp)$, can only carry a current of $dI = 2e(\partial E/\partial p)(dp/h) = (2e/h)dE$, where $\partial E/\partial p$ is the velocity of the electron at the Fermi level and dp/h is the number of electrons, irrespective of the internal structure of the chain. Second, the number of allowable conduction channels, determined by the electron motion normal to the strand axis, is one for a single strand. The double strand, then, has twice the unit conductance if the interaction between the two individual rows is not strong.

The current atom bridges appear to differ from the QPCs in STM configurations^{4–7}. For the QPCs, it is claimed that only narrow,

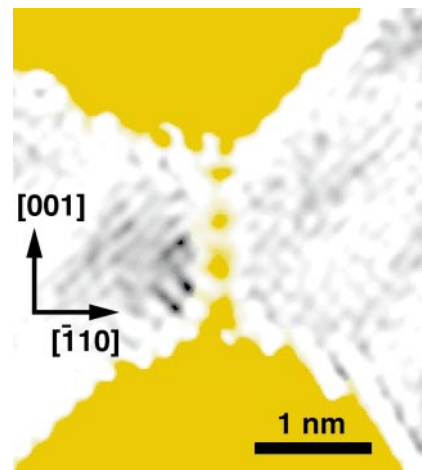


Figure 4 Electron microscope image of a linear strand of gold atoms (four coloured dots) forming a bridge between two gold films (coloured areas). The spacings of the four gold atoms are $0.35\text{--}0.40 \text{ nm}$. The strand is oriented along the $[001]$ direction of the gold (110) film. This image was processed to highlight the linear strand, where the lattice fringes of the gold film in the original electron microscope image were filtered out by Fourier transform.

nanometre-wide necks form between tip and substrate. Molecular dynamics simulations^{5,6,21,22} predict a narrow neck, but we did not find this here. A recent molecular dynamics simulation²³ reports elongated gold $[110]$ and $[001]$ wire structures in relation with the quantized conductance: the gold $[110]$ wire (Fig. 8c of ref. 23) becomes permanently disordered, however; the $[001]$ wire in Fig. 7c of ref. 23 is similar to the single atomic strand shown in our Fig. 4. This strand, however, has much larger atom spacings than those indicated by the molecular dynamics simulation. Regardless of the differences, some interesting issues remain to be investigated: the relation between the number of atomic rows in the strand and the number of channels for conducting electrons, quantum fluctuations of conductances that have been reported for metallic QPCs⁷ and were also observed in the experiments described here following breakage of the strands, and the elementary excitation of conduction electrons in the strands.

The current linear strands persist for as long as two minutes. The gold strands in these experiments were stable, regular and uniform to an extent previously unrealized. The stability, regularity and uniformity of these strands of gold atoms are thought to be inherent to their natural, self-ordering phenomena, and may allow applications in devices as digitized units of electron conductors. □

Received 4 March; accepted 29 June 1998.

- Webb, R. A., Washburn, S., Umbach, C. P. & Laibowitz, R. B. Observation of h/e Aharonov–Bohm oscillations in normal-metal rings. *Phys. Rev. Lett.* **54**, 2696–2699 (1985).
- Iijima, S. Helical microtubules of graphitic carbon. *Nature* **354**, 56–58 (1991).
- Bockrath, M. *et al.* Single-electron transport in ropes of carbon nanotubes. *Science* **275**, 1922–1925 (1997).
- Agrait, N., Rodrigo, J. G. & Vieira, S. Conductance steps and quantization in atomic-size contacts. *Phys. Rev. B* **47**, 12345–12348 (1993).
- Pascual, J. I. *et al.* Properties of metallic nanowires: from conductance quantization to localization. *Science* **267**, 1793–1795 (1995).
- Olesen, L. *et al.* Quantized conductance in an atom-size point contact. *Phys. Rev. Lett.* **72**, 2251–2254 (1994).
- Costa-Krämer, J. L. *et al.* Conductance quantization in nanowires formed between micro- and macroscopic metallic electrodes. *Phys. Rev. B* **55**, 5416–5424 (1997).
- Müller, C. J., Krans, J. M., Todorov, T. N. & Reed, M. A. Quantization effects in the conductance of metallic contacts at room temperature. *Phys. Rev. B* **53**, 1022–1025 (1996).
- Landman, U., Luedtke, W. D., Salisburry, B. E. & Whetten, R. L. Reversible manipulations of room-temperature mechanical and quantum transport properties in nanowire junctions. *Phys. Rev. Lett.* **77**, 1362–1365 (1996).
- Costa-Krämer, J. L., García, N., García-Mochales, P. & Serena, P. A. Nanowire formation in macroscopic metallic contacts: quantum mechanical conductance tapping a table top. *Surf. Sci.* **342**, L1144–L1149 (1995).
- Hansen, K., Lægsgaard, E., Stensgaard, I. & Besenbacher, F. Quantized conductance in relays. *Phys. Rev. B* **56**, 2208–2220 (1997).
- Yasuda, H. & Sakai, A. Conductance of atomic-scale gold contacts under high-bias voltages. *Phys. Rev. B* **56**, 1069–1072 (1997).
- van Wees, B. J. *et al.* Quantized conductance of point contacts in a two-dimensional electron gas. *Phys. Rev. Lett.* **60**, 848–850 (1988).

14. Wharam, D. A. *et al.* One-dimensional transport and the quantisation of the ballistic resistance. *J. Phys. C* **21**, L209–L214 (1988).
15. Cowley, J. M. in *Diffraction Physics* 62–63 (Elsevier, Amsterdam, 1975).
16. Kondo, Y. & Takayanagi, K. Gold nanobridge stabilized by surface structure. *Phys. Rev. Lett.* **79**, 3455–3458 (1997).
17. Maxwell, J. C. in *A Treatise on Electricity and Magnetism* (Clarendon, Oxford, 1904).
18. Sharvin, Y. V. A possible method for studying fermi surfaces. *Sov. Phys. JETP* **21**, 655–656 (1965).
19. Landauer, R. Spatial variation of currents and fields due to localized scatterers in metallic conduction. *IBM J. Res. Dev.* **1**, 223–231 (1957).
20. Kubo, R. Statistical-mechanical theory of irreversible process. I. General theory and simple applications to magnetic and conduction problems. *J. Phys. Soc. Jpn* **12**, 570–586 (1957).
21. Landman, U., Luedtke, W. D., Burnham, N. A. & Colton, R. J. Atomic mechanics and dynamics of adhesion, nanoindentation, and fracture. *Science* **248**, 454–461 (1990).
22. Bratkovsky, A. M., Sutton, A. P. & Todorov, T. N. Conditions for conductance quantization in realistic models of atomic-scale metallic contacts. *Phys. Rev. B* **52**, 5036–5051 (1995).
23. Sørensen, M. R., Brandbyge, M. & Jacobsen, K. W. Mechanical deformation of atomic-scale metallic contacts: structure and mechanisms. *Phys. Rev. B* **57**, 3283–3294 (1998).

Correspondence and requests for materials should be addressed to H.O. (e-mail: ohnishi@tapro.jst.go.jp).

Formation and manipulation of a metallic wire of single gold atoms

A. I. Yanson*, G. Rubio Bollinger†, H. E. van den Brom*, N. Agrait† & J. M. van Ruitenbeek*

*Kamerlingh Onnes Laboratorium, Leiden University, PO Box 9504, NL-2300 RA Leiden, The Netherlands

†Laboratorio de Bajas Temperaturas, Dept Física de la Materia Condensada C-III, Instituto Universitario de Ciencia de Materiales “Nicolás Cabrera”, Universidad Autónoma de Madrid, E-28049 Madrid, Spain

The continuing miniaturization of microelectronics raises the prospect of nanometre-scale devices with mechanical and electrical properties that are qualitatively different from those at larger dimensions. The investigation of these properties, and particularly the increasing influence of quantum effects on electron transport, has therefore attracted much interest. Quantum properties of the conductance can be observed when ‘breaking’ a metallic contact: as two metal electrodes in contact with each other are slowly retracted, the contact area undergoes structural rearrangements until it consists in its final stages of only a few bridging atoms^{1–3}. Just before the abrupt transition to tunnelling occurs, the electrical conductance through a monovalent metal contact is always close to a value of $2e^2/h$ ($\approx 12.9 \text{ k}\Omega^{-1}$), where e is the charge on an electron and h is Planck’s constant^{4–6}. This value corresponds to one quantum unit of conductance, thus indicating that the ‘neck’ of the contact consists of a single atom⁷. In contrast to previous observations of only single-atom necks, here we describe the breaking of atomic-scale gold contacts, which leads to the formation of gold chains one atom thick and at least four atoms long. Once we start to pull out a chain, the conductance never exceeds $2e^2/h$, confirming that it acts as a one-dimensional quantized nanowire. Given their high stability and the ability to support ballistic electron transport, these structures seem well suited for the investigation of atomic-scale electronics.

The experimental techniques used for studies on metallic contacts of atomic dimensions are all based on piezoelectric transducers which allow fine positioning of two metal electrodes with respect to each other. Scanning tunnelling microscopy (STM), in which the tip is driven into contact with a metal surface and the conductance is measured during subsequent retraction, has been widely used for this purpose^{4,5,8,9}. The other commonly used technique employs a mechanically controllable break-junction (MCB). In this case, one starts with a macroscopic notched wire¹⁰, or a nanofabricated metal bridge¹¹ mounted on a flexible substrate. The wire (or bridge) is broken at low temperatures in vacuum, and contact is re-established between the fracture surfaces by piezoelectric control of substrate

bending. Here we have used both MCB and a very stable STM at liquid-helium temperatures to produce and study chains of single gold atoms. In each case, high-purity (>99.99%) gold was used. Conductance was measured at a 10 mV d.c. voltage bias with 1% accuracy.

An example of a conductance curve obtained while stretching a gold nanocontact in an MCB is presented in Fig. 1. The curve reflects the evolution of some particular atomic configuration, during which the conductance decreases in a series of sharp vertically descending steps, with a gradual slope on the plateaux in between. These steps have been shown to be the result of atomic structural rearrangements^{9,12}. The curves for successive rupture sequences do not repeat in detail, as they depend on the exact atomic positions in the contact. However, many curves have one striking feature in common: the remarkable length of the last conductance plateau just before rupture, at the value of about one conductance quantum, $G_0 = 2e^2/h$. Previous experiments have shown that the conductance at the last plateau for monovalent metals is usually close to $1 G_0$ (refs 4–6), and it has been argued that these plateaux correspond to a contact with a single atom at the narrowest cross-section⁷. Examples have been found previously of exceptionally long stretched last plateaux, in particular for gold (see, for example, ref. 20), but this has not received due attention. The interest becomes clear when we recognize that the conductance of a point contact is determined predominantly by the size of its narrowest cross-section and that, during contact elongation, all structural transformations are localized to the neck region^{2,3,12}. Figure 1 shows that during the last stage of elongation the conductance stays in a limited range of values corresponding to one atom in cross-section, while the contact is being stretched over distances up to 20 Å. This suggests the possibility that the contact stretches to form a chain of single atoms. As we cannot image the

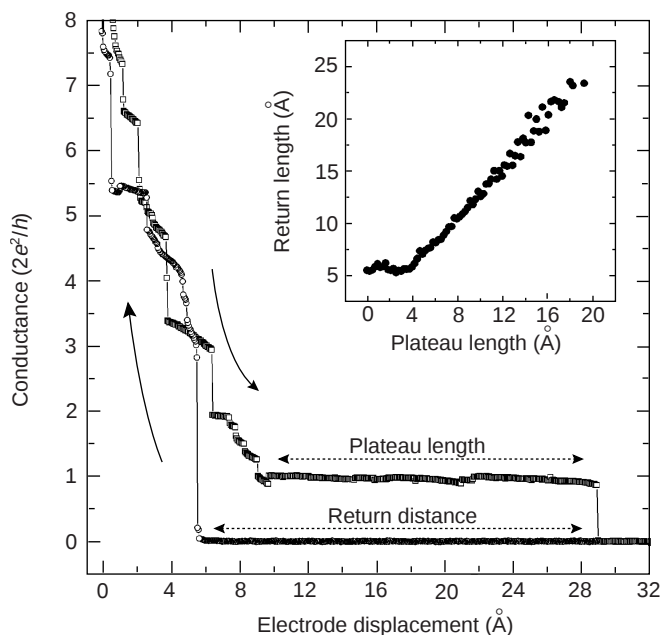


Figure 1 Conductance as a function of the displacement of the two gold electrodes with respect to each other in an MCB experiment at 4.2 K. The trace starts at the upper left, coming from higher conductance values (open squares). A long plateau with a conductance near $1 G_0$ is observed and, after a jump to tunnelling, the electrode needs to return by a little more than the length of the long plateau to come back into contact (open circles). Inset, the average of the return distance as a function of the length of the long plateau (recorded with an STM at 4.2 K). The relation is approximately 1:1, with an offset of 5 Å which is probably due to the elasticity of the atomic structure.

atomic structure of the neck directly, we performed the following experiments in order to test this hypothesis.

Figure 1 illustrates that the distance the electrode needs to travel back to re-establish contact after rupture is almost equal to the length of the last plateau itself. We have recorded these lengths for a large number of contacts, and the average return distance is plotted in Fig. 1 inset as a function of the length of the last plateau. The relation is approximately 1:1, suggesting that a fragile structure is formed with a length corresponding to that of the last plateau, which is unable to support itself when it breaks and collapses onto the 'banks' on either side.

The probability distribution for pulling a long last plateau as a function of its length was obtained by recording a histogram of plateau lengths (Fig. 2). Instead of a smooth distribution, we find a series of three equidistant peaks, with a shoulder at the fourth and fifth position. The probability of pulling a structure of length L decreases rapidly for large L and drops below 10^{-4} for $L > 20$ Å, so the longest observed plateaux have a length of 25 Å. The shape of the histogram suggests that the nanobridges tend to be elongated by integer multiples of 3.6 Å, which is somewhat longer than the nearest-neighbour spacing 2.88 Å of gold atoms in the crystal.

As a further test, we carried out STM experiments during which one of the electrodes was moved laterally once a plateau of a given length was formed. For a short length of the last plateau (Fig. 3a), when we start oscillating the STM tip sideways (Fig. 3b), we observe rupture of the contact at low oscillation amplitudes. Repeating the experiment for longer elongation of the last plateau (Fig. 3c), fracture is observed at much larger lateral displacement amplitudes (Fig. 3d) and the bridge is found to collapse. The small jumps in the conductance in Fig. 3b and d have the same origin as those found during elongation of the last plateau in Fig. 1. That is, they are due to atomic structural rearrangements, which have only a minor effect on the conductance as long as the narrowest cross-section is still a single atom. The two atomic configurations at each side of the small conductance jump are separated by an energy barrier, which can be surmounted by applying a sufficiently large force. This force stretches the atomic bridge over a length proportional to its elasticity. Consequently, the hysteresis of the conductance jumps that occur as the structure is moved laterally (Fig. 3d) indicates that longer structures have a larger lateral elasticity than short ones.

We find that there is a significant probability for formation of a long bridge with a length given by that of the last plateau, which

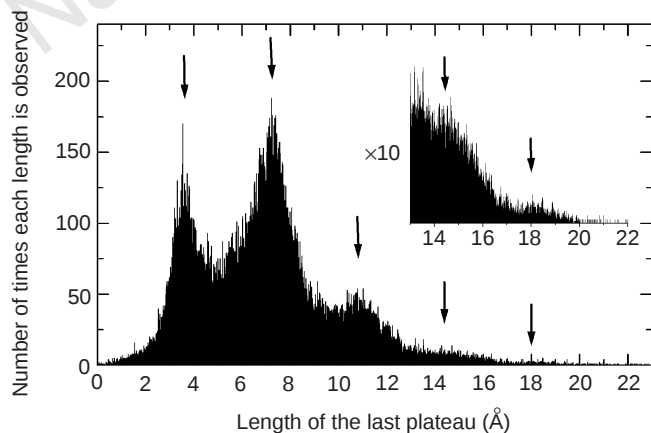


Figure 2 The distribution of lengths for the last plateau, obtained in ~100,000 experiments similar to those described in Fig. 1, shows a number of equidistant maxima. The arrows are positioned at multiples of 3.6 Å. The data were recorded with an MCB at 4.2 K. The length of the last plateau was defined as the distance between the points at which the conductance drops below $1.2G_0$ and $0.8G_0$, respectively. Inset, the tail of the distribution on a 10x expanded scale. The accuracy for the calibration of the length scale is 30%.

completely collapses on breaking. Its smallest cross-section is one atom and it breaks at multiples of ~3.6 Å in length, which would correspond to a stretched Au–Au bond distance. We can swing the bridge sideways by a distance comparable to its length. We conclude that all the evidence combines to show that we are pulling, atom by atom, a freely suspended chain of single Au atoms. The peaks in the plateau length distribution shown in Fig. 2 correspond to the observation of chains containing up to 4 or 5 atoms. Although we cannot resolve the peak structure for longer plateaux, extrapolating the periodic structure suggests that the longest plateaux observed correspond to chains of seven atoms.

Despite its low probability of formation, once an atomic chain is pulled it remains very stable: some of the longest chains obtained in our experiments have been held stable for as long as 1 hour, after which we stopped the experiment. They can sustain very large current densities of up to $8 \times 10^{14} \text{ A m}^{-2}$ (currents up to 80 μA or a voltage up to 1 V), proving that the electron transport is ballistic, and that most of the power is dissipated in the electrodes far away from the contact. This makes the atomic chains suitable for investigation as conductors in atomic electronic circuits.

Once the conductance drops below $1 G_0$ at the beginning of chain formation, it never rises above this value. This is consistent with the

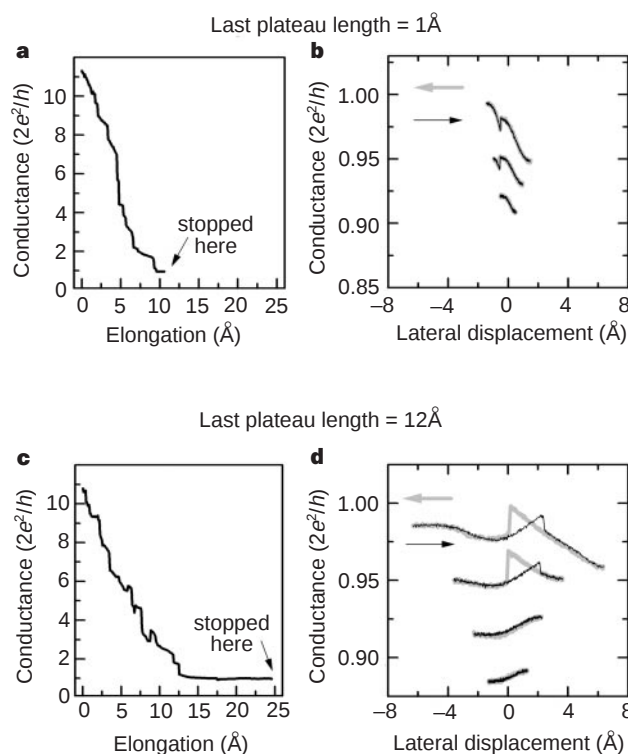


Figure 3 Swinging an atomic contact sideways by lateral displacement of one of its ends in an STM experiment. First, the elongation of the last plateau is stopped early (a), at a length of ~1 Å (marked with an arrow). Then, one end of the contact is cyclically displaced perpendicular to the contact axis (b) while slowly increasing the displacement amplitude. The curves for smaller swing amplitudes are shifted by increments of $-0.03 G_0$ for clarity. The contact breaks at a lateral displacement of 1.4 Å. Performing the same experiment for a last plateau of 12 Å length (c) gives quite different results: rupture occurs at much larger lateral swing amplitudes of 6.3 Å (d) and the contact bridge collapses. For small swing amplitudes the contact behaves elastically and the conductance for both directions of the swing superpose. As the amplitude is increased, small hysteric jumps in conductance take place (note the expanded scale), resulting from slight rearrangements of the atoms in the chain or at its bases.

notion of the conductance being perfectly described by one single quantum mode⁷. Fluctuations to lower values correspond to a reduced transmission probability for this mode as a result of back-scattering, but the maximum value should be limited to $1 G_0$, in agreement with the observations. In this sense, our metallic atomic wire is a perfectly quantized one-dimensional conductor. Previously, Yazdani *et al.*¹³ succeeded in measuring the conductance of a chain of two atoms of the noble-gas xenon, which was found to be orders of magnitude lower than the quantum unit of conductance due to the lack of metallic bonding. A recent calculation by Lang¹⁴ on single atomic chains of up to four sodium atoms predicts their conductance to fall below $0.7 G_0$, fluctuating with the chain length. This is somewhat smaller than the experimental values, which almost all fall in the range from 0.8 to $1 G_0$ (see, for example, Figs 1 and 3). This discrepancy may be influenced by the interface to the 'jellium' electrodes used in the calculation.

Recent molecular-dynamics simulations of metallic nanocontacts by Sutton and Todorov (unpublished work), Sørensen *et al.*^{15,16} and Finbow *et al.*¹⁷ support the plausibility of occasional formation of atomic chains, although these authors note that their model interatomic potential may not be reliable for these unusual one-dimensional structures. Advanced first-principles molecular-dynamics calculations have been performed for Na nanocontacts at 190 K (ref. 18), and showed no indication of chain formation. Although it is not clear to us what determines the difference in behaviour between gold and sodium, this result is consistent with the fact that we do not observe chain formation for sodium in the low-temperature experiments.

The atomic-chain structures should open many avenues of further investigations. Studying the mechanical properties of this unusual form of matter should enable stringent tests of our understanding of interatomic potentials. One might search for evidence of one-dimensional excitations, such as phonons and plasmons. Further, the electronic properties of these chains are expected to be those of a perfect one-dimensional conductor, where the electron-phonon interaction might lead to a Peierls transition. The electron-electron interaction may lead to the formation of a so-called Tomonaga-Luttinger liquid¹⁹, which replaces the familiar Fermi liquid description for bulk metals when the conductor becomes one-dimensional. □

Received 2 March; accepted 11 September 1998.

1. van Ruitenbeek, J. M. in *Mesoscopic Electron Transport* (eds Sohn, L. L., Kouwenhoven, L. P. & Schön, G.) 549–579 (NATO ASI Ser. E, Vol. 345, Kluwer Academic, Dordrecht, 1997).
2. Sutton, A. P. & Pethica, J. B. Inelastic flow processes in nanometre volumes of solids. *J. Phys.: Condens. Matter* **2**, 5317–5326 (1990).
3. Landman, U., Luedtke, W. D., Burnham, N. A. & Colton, R. J. Atomistic mechanisms and dynamics of adhesion, nanoindentation and fracture. *Science* **248**, 454–461 (1990).
4. Agraït, N., Rodrigo, J. G. & Vieira, S. Conductance steps and quantization in atomic-size contacts. *Phys. Rev. B* **47**, 12345–12348 (1993).
5. Pascual, J. I. *et al.* Quantum contact in gold nanostructures by scanning tunneling microscopy. *Phys. Rev. Lett.* **71**, 1852–1855 (1993).
6. Krans, J. M. *et al.* One-atom point contacts. *Phys. Rev. B* **48**, 14721–14724 (1993).
7. Scheer, E. *et al.* The signature of chemical valence in the electrical conduction through a single-atom contact. *Nature* **394**, 154–157 (1998).
8. Brandbyge, M. *et al.* Quantized conductance in atom-sized wires between two metals. *Phys. Rev. B* **52**, 8499–8514 (1995).
9. Rubio, G., Agraït, N. & Vieira, S. Atomic-sized metallic contacts: mechanical properties and electronic transport. *Phys. Rev. Lett.* **76**, 2302–2305 (1996).
10. Muller, C. J., van Ruitenbeek, J. M. & de Jongh, L. J. Experimental observation of the transition from weak link to tunnel junction. *Physica C* **191**, 485–504 (1992).
11. van Ruitenbeek, J. M. *et al.* Adjustable nanofabricated atom size contacts. *Rev. Sci. Instrum.* **67**, 108–111 (1996).
12. Todorov, T. N. & Sutton, A. P. Jumps in electronic conductance due to mechanical instabilities. *Phys. Rev. Lett.* **70**, 2138–2141 (1993).
13. Yazdani, A., Eigler, D. M. & Lang, N. D. Off-resonance conduction through atomic wires. *Science* **272**, 1921–1924 (1996).
14. Lang, N. D. Anomalous dependence of resistance on length in atomic wires. *Phys. Rev. Lett.* **79**, 1357–1360 (1997).
15. Brandbyge, M., Sørensen, M. R. & Jacobsen, K. W. Conductance eigenchannels in nanocontacts. *Phys. Rev. B* **56**, 14956–14959 (1997).
16. Sørensen, M. R., Brandbyge, M. & Jacobsen, K. W. Mechanical deformation of atomic-scale metallic contacts: structure and mechanism. *Phys. Rev. B* **57**, 3283–3295 (1998).
17. Finbow, G. M., Lynden-Bell, R. M. & McDonald, I. R. Atomic simulation of the stretching of nanoscale metal wires. *Mol. Phys.* **92** (N4), 705–714 (1997).
18. Barnett, R. N. & Landman, U. Cluster-derived structures and conductance fluctuations in nanowires. *Nature* **387**, 788–791 (1997).
19. Fisher, M. P. A. & Glazman, L. I. in *Mesoscopic Electron Transport* (eds Sohn, L. L., Kouwenhoven, L. P. & Schön, G.) 331–373 (NATO ASI Ser. E, Vol. 345, Kluwer Academic, Dordrecht, 1997).
20. Krans, J. M. *Size Effects in Atomic-Scale Point Contacts*. Thesis, Leiden Univ. (1996).

Acknowledgements. We thank B. Ludoph for many discussions, S. Vieira and L. J. de Jongh for discussions and continuous support, and A. P. Sutton, T. N. Todorov, M. R. Sørensen, M. Brandbyge and K. W. Jacobsen for communicating their results before publication. A.I.Y., H.E.v.d.B. and J.M.v.R. were supported by FOM; N.A. and G.R.B. were supported by the CICYT.

Correspondence and requests for materials should be addressed to J.M.v.R. (e-mail: ruitenbe@ruikol. LeidenUniv.nl).

Contribution of bedrock nitrogen to high nitrate concentrations in stream water

J. M. Holloway*, R. A. Dahlgren*, B. Hansen† & W. H. Casey*

* Department of Land, Air and Water Resources, University of California, Davis, California 95616, USA

† Department of Earth Sciences, University of Aarhus, Ny Munkegade, Building 520, DK-800 Aarhus C, Denmark

Concentrations of nitrate in stream water throughout the world are reported to be elevated relative to natural background levels. This enrichment is commonly attributed to anthropogenic activities such as atmospheric emissions¹, livestock feeding², agricultural runoff^{3,4}, timber harvesting practices⁵ and domestic/industrial effluent discharge^{4,6}. Here we show that bedrock containing appreciable concentrations of fixed nitrogen contribute a surprisingly large amount of nitrate to surface waters in certain California watersheds, to an extent that even small areas of these rocks have a profound influence on water quality. As 75% of the rocks now exposed at the Earth's surface are sedimentary in origin⁷, and as these rocks contain about 20% of the global nitrogen inventory⁸, 'geological' nitrogen may be a large and hitherto unappreciated source of nitrate to surface waters. Such a natural nitrate source may be especially significant given that nitrate contamination at very low levels can contribute to surface water eutrophication⁹, may cause infant methaemoglobinemia ('blue baby' syndrome)⁶ and has been implicated in certain cancers⁶. In addition, geological nitrogen may be a source of the 'missing' nitrogen noted in several biogeochemical studies of ecosystem nitrogen budgets¹.

The source(s) of nitrate contributing to eutrophication and periodic fish kills in downstream reservoirs of the Mokelumne River watershed in the central Sierra Nevada of California has been an enigma. We monitored water quality throughout the watershed over the past 2.5 years to identify the primary source(s) of nitrate entering the downstream reservoirs. Due to the close relationship we observed between streamwater nitrate concentrations and bedrock nitrogen concentrations, we specifically examined whether the bedrock could be a source of streamwater nitrate.

The Mokelumne River watershed covers an area of 980 km² extending from the Central Valley (50 m elevation) to the crest of the Sierra Nevada (3,300 m) (Fig. 1). The upper part of the watershed is dominated by Mesozoic granite and diorite and Tertiary volcanic mudflows while Palaeozoic and Jurassic rocks lie to the west in a belt of metasedimentary and metavolcanic lithologies flanking the Sierra Nevada¹⁰. Streamwater nitrate concentrations in the upper elevation watersheds have median values less than 2 µM (Fig. 1). Maximum values (~10 µM) occur during storm events in the late autumn–early winter, and low concentrations (<0.4 µM) are associated with the large amount of runoff (~70%) originating from the melting snowpack. In contrast, tributary streams in the lower watershed generally have elevated nitrate

concentrations with median and maximum nitrate concentrations of 18–99 μM and 116–625 μM , respectively. These streams lie in a geological region dominated by metasedimentary and metavolcanic rocks^{11,12}.

Because of the close association of high-nitrate streams and the metasedimentary and metavolcanic bedrock, we determined nitrogen concentrations in several rocks to determine whether the bedrock could be a possible source of nitrogen. Concentrations of total nitrogen¹³ were highest in phyllite ($1,034 \pm 255$ mean $\pm$ standard error of the mean, range 648–1,321 mg N kg^{-1} , $n = 5$) followed by slate (997 ± 377 , range 345–1,801 mg N kg^{-1} , $n = 19$), biotite schist (717 ± 145 , range 253–1,321 mg N kg^{-1} , $n = 8$), metavolcanic breccia (385 ± 92 , range 98–739 mg N kg^{-1} , $n = 10$), and greenstone (245 ± 52 , range 108–430 mg N kg^{-1} , $n = 8$). Nitrogen occurs in both organic and inorganic forms in the slate, whereas it is dominantly in inorganic forms in the other lithologies. The inorganic nitrogen is assumed to be present as NH_4^+ incorporated into the interlayer of muscovite and sericite; these minerals are prevalent in all lithologies. When nitrogen concentrations are expressed on a landscape area basis, a 10-cm thickness of phyllite, slate, biotite schist, and metavolcanic breccia contains 2,688, 2,592, 1,807, and 1,001 kg N ha^{-1} , respectively, based on the average total nitrogen concentration and rock density. Thus, the bedrock contains an appreciable pool of nitrogen that could be released upon weathering. In comparison, total nitrogen concentrations in intrusive igneous rocks (for example, granite and diorite) from the higher elevations of the watershed were generally $<50 \text{ mg N kg}^{-1}$.

Soils formed on slate are weakly developed, with a thin A horizon (10 cm) overlying weakly altered C horizon (30 cm) and saprolite (10 cm). Total nitrogen was determined for rock, saprolite, and the sand, silt and clay fractions of the A and C horizons after removal of soil organic matter, to determine how much nitrogen was released during weathering of slate. Nitrogen was lost from the saprolite, sand and silt fractions but not from the clay fraction (Table 1). The lack of depletion in the clay fraction may result from fixation¹⁴ of some NH_4^+ in the interlayer position of the vermiculite and smectite that dominate the clay fraction. A mass balance calculated for the alteration of rock to soil indicates a loss of 2,400 kg N ha^{-1} for the

Table 1 Total nitrogen concentrations in rock, saprolite and soil

Component	Total N*, mean ($\pm$ s.e.m.) (mg kg^{-1})	Particle-size analysis (%)
Rock	1,447 (106) a†	
Saprolite	1,229 (88) b	
C horizon		
Sand	1,121 (89) bc	60
Silt	950 (20) c	33
Clay	1,419 (143) a	7
A horizon		
Sand	978 (42) c	63
Silt	983 (38) c	29
Clay	1,460 (14) a	8

* Rock and soil nitrogen concentrations ($n = 5$) were determined by combustion at $1,020^\circ\text{C}$ using a Carlo Erba elemental analyser¹³. All rock, saprolite and soil samples (that is, sand, silt and clay) were treated with 5% hydrogen peroxide to remove soil organic matter before determination of nitrogen concentrations. Rock and saprolite samples were powdered before hydrogen peroxide treatment and treated identically to soil samples.

† Mean values in column followed by the same lower-case letter are not significantly different ($p = 0.05$) as determined by analysis of variance with Fisher's least significance difference test for pair-wise comparisons. All statistical analysis was performed using SYSTAT³².

soil profile, showing that much nitrogen has been released during the rock-to-soil transformation.

Within the lower watershed, we identified adjacent watersheds having biotite schist (Mica and Rattlesnake) and diorite (Electra) bedrock lithologies (Fig. 2). Mean nitrogen concentrations in the biotite schist and diorite were 717 and $<50 \text{ mg N kg}^{-1}$, respectively. With the exception of bedrock lithology, the Mica and Electra watersheds have identical hydrological response (ephemeral), land-use, vegetation, microclimate and topography. Rattlesnake watershed has similar factors to Mica and Electra except that it is a larger watershed and has a longer duration of flow into the summer. Nitrate concentrations in the watershed with diorite lithology (Electra) remain consistently low (median 3.3 μM) with many concentrations undetectable ($<0.4 \mu\text{M}$) (Fig. 2). In sharp contrast, streams in the watersheds having biotite schist lithologies (Mica and Rattlesnake) show a pattern of maximum nitrate concentrations ($>300 \mu\text{M}$) during the beginning of the rainy season (December–January) and decreasing concentrations during the

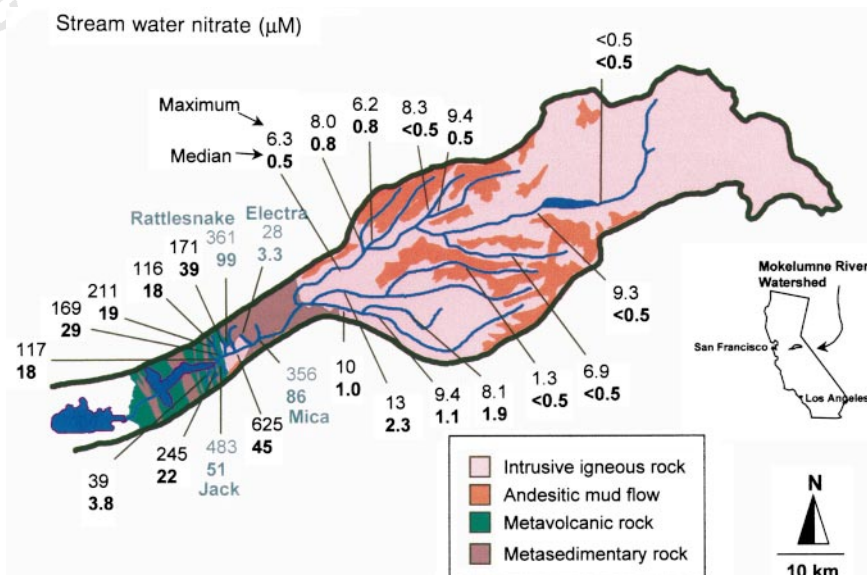


Figure 1 Surficial geology of the Mokelumne River watershed. Streamwater nitrate concentrations (median and maximum) are shown for the period January 1995 to September 1997. The geological map, modified from previous workers^{10–12}, includes four main groups of rock: (1) intrusive igneous rock (primarily diorite and granite), (2) andesitic mud flows, (3) metavolcanic rock (basalt metamorphosed to

greenstone) and (4) metasedimentary rock (dominantly slate, phyllite and schist). Streamwater samples were grab samples collected biweekly during the rainy season (December to May) and monthly during the dry season. Highlighted values with names (Jack, Rattlesnake, Mica and Electra) are referred to in text. Streamwater nitrate concentrations were determined by ion chromatography.

spring period before the cessation of streamflow. This paired watershed comparison provides strong evidence that differences in streamwater nitrate concentrations are attributable to differences in bedrock lithology rather than other watershed factors.

All streams in the lower portion of the Mokelumne River watershed have a rapid hydrograph response, with a short lag time from precipitation to peak discharge and low discharge ($<0.2 \text{ m}^3 \text{ s}^{-1}$) during baseflow (Fig. 3). This hydrological response suggests that most of the streamflow during storm events originates from interflow through the vadose-zone (that is, unsaturated soil zone) with little contribution from ground water^{15,16}. Streamwater nitrate concentrations increase to maximum levels during the first storm events of the water year (October to September time period). In subsequent storms, nitrate concentrations decrease during the peak of the hydrograph due to dilution. There is a progressive decrease in nitrate concentrations with each storm event during the water year.

To elucidate the mechanism responsible for this temporal pattern in streamwater nitrate concentrations, the concentrations of mineral nitrogen ($\text{NH}_4^+ + \text{NO}_3^-$) in the soil profile were determined. Soil samples were collected from the C horizon of a watershed with slate lithology; soil solutions were extracted by centrifugation¹⁷, and mineral nitrogen was extracted with 1 M KCl (ref. 18). The C horizon was selected because it contains very low concentrations of soil organic matter ($<2 \text{ g C per kg soil}$) and therefore mineral nitrogen concentrations are assumed to originate primarily from the mineral soil fraction. Mineral nitrogen concentrations of *in situ* soil solutions and KCl extracts show a large decrease between the initial wetting of the soil profile and subsequent sampling periods (Fig. 4). We interpret these data to indicate that mineral nitrogen is concentrated in the soil micropores (that is, the soil solution extracted by centrifugation)¹⁹ during the summer–autumn period when there is limited rainfall. The mineral nitrogen is progressively leached from the micropores throughout the rainy season, leading to maximum streamwater nitrate concentrations in the early winter (December–January) and minimum concentrations before cessation of streamflow in the spring.

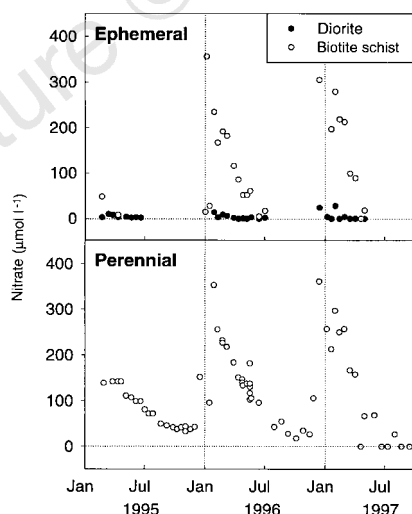


Figure 2 Streamwater nitrate concentrations. Top panel, data for paired ephemeral streams, one with diorite bedrock (Electra) and the other with biotite schist (Mica). These watersheds have identical land use (range land), vegetation (oak savannah with annual grasses), topography (south aspect; 6–24% slopes) and watershed area (~20 ha). Bottom panel, data from Rattlesnake Creek watershed (~100 ha), a perennial stream draining a nearby biotite-schist watershed having identical vegetation and land use to Mica and Electra watersheds. Streams in the lower watershed are dominantly ephemeral, flowing only during the high-precipitation periods (late December to February, with sporadic rain events to the end of May).

On the basis of simultaneous measurements of stream discharge and nitrate concentrations in a watershed with phyllite lithology (Jack Creek), we calculate a nitrogen flux of $19.9 \text{ kg N ha}^{-1} \text{ yr}^{-1}$ during the 1995–96 water year. The nitrogen flux for the upper Mokelumne watershed, dominated by intrusive igneous rocks and andesitic mudflow, was measured below the confluence of the three main tributaries and showed only $0.12 \text{ kg N ha}^{-1} \text{ yr}^{-1}$. An estimated mass balance for the entire watershed suggests that greater than 90% of the nitrate originates from the lower watershed that contains geological nitrogen (the upper watershed has 90% of the watershed area and a nitrogen flux of $0.12 \text{ kg N ha}^{-1} \text{ yr}^{-1}$; the lower watershed has 10% of the watershed area with nitrogen fluxes of 10–20 $\text{kg N ha}^{-1} \text{ yr}^{-1}$). Thus the watershed area that contains geological nitrogen contributes a disproportionately large amount of the total nitrate to the downstream reservoirs.

We believe, on the basis of our investigation, that release of geological nitrogen contributes to nitrogen saturation²⁰ of these ecosystems, leading to elevated streamwater nitrate concentrations. We measured additions of about $2 \text{ kg N ha}^{-1} \text{ yr}^{-1}$ in bulk precipitation samples at both low- and high-elevation sites in the Mokelumne River watershed. However, this atmospheric deposition flux by itself can not explain the nearly $20 \text{ kg N ha}^{-1} \text{ yr}^{-1}$ flux in stream water. Historically, oak trees probably attenuated the deep leaching of nitrate by cycling nitrogen back to the soil surface as litter-fall²¹ where periodic wildfires (at a frequency of one per 30–50 years; ref. 22) volatilize nitrogen⁸. Thinning of oak trees for pasture improvement²¹ and fire suppression have probably exacerbated nitrogen saturation in these ecosystems so that the adverse effects of geological nitrogen on water quality can only now be recognized.

Elevated concentrations of nitrate in surface waters are reported throughout the world, and are commonly attributed to anthropogenic activities^{1–6}. The results reported here are, to our knowledge, the first demonstration of the role of geological nitrogen as a non-point source of nitrate contamination to surface waters. The same geological formations that contain nitrogen in the Mokelumne River watershed extend for 300 km along the western flank of the Sierra Nevada, indicating the potential for nitrate contamination of

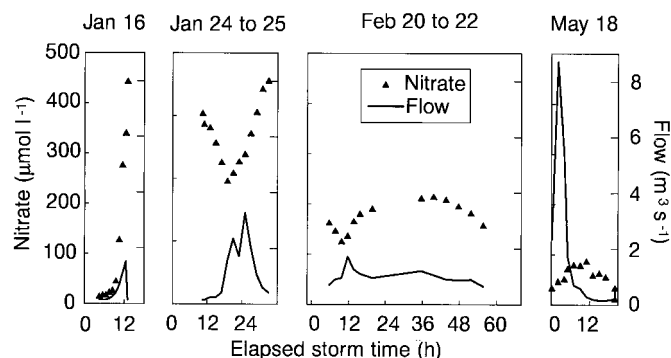


Figure 3 Stream discharge and nitrate concentrations during storm events. Data are for Jack Creek watershed (phyllite lithology) for the 1995–96 water year. Time elapsed from the onset of storm events is shown in hours on the abscissa. A stream autosampler was used to collect samples during storm events. Stream discharge was calculated from continuous stage measurements recorded with a pressure transducer and data logger. Streams show a rapid response to storm events and return quickly to low baseflow levels after cessation of rainfall.

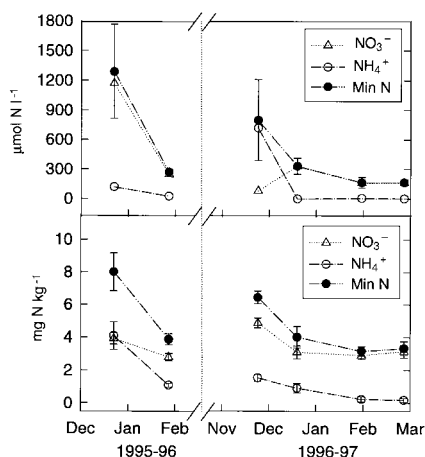


Figure 4 Mineral nitrogen ($\text{NO}_3^- + \text{NH}_4^+$) concentrations for C horizon soils. These soils have phyllite parent rock in Jack Creek watershed. Top panel, data for *in situ* soil solutions extracted from soil samples via centrifugation¹⁷, bottom panel, KCl-extractable nitrogen¹⁸ from the same soil samples ($n = 3$). Data from the 1995–96 and 1996–97 water years show a similar pattern of high total mineral N at the beginning of the water year, and lower concentrations following progressive leaching of the soil profile.

much of California's surface water supply. Given the widespread occurrence of sedimentary and metasedimentary rocks on the Earth's surface^{7,23–31} and the large inventory of nitrogen contained within these rocks⁸, the role of geological nitrogen as a non-point source of nitrate contamination in surface waters needs to be re-evaluated. □

Received 7 July 1997; accepted 28 July 1998

- Galloway, J. N., Schlesinger, W. H., Levy, H., Michaels, A. & Schnoor, J. L. Nitrogen fixation: anthropogenic enhancement—environmental response. *Glob. Biogeochem. Cycles* **9**, 235–252 (1995).
- Power, J. F. & Papendick, R. I. in *Fertilizer Technology and Use* (ed. Engelstad, O. P.) 503–520 (Soil Sci. Soc. Am., Madison, Wisconsin, 1985).
- Spalding, R. F. & Exner, M. E. Occurrence of nitrate in groundwater—a review. *J. Environ. Qual.* **22**, 393–402 (1993).
- Keeney, D. R. in *CRC Critical Reviews in Environmental Control* (ed. Straub, C. P.) 257–304 (CRC, Boca Raton, Florida, 1986).
- Likens, G. E., Bormann, F. H., Johnson, N. M., Fisher, D. W. & Pierce, R. S. Effects of forest cutting and herbicide treatment on nutrient budgets in the Hubbard Brook watershed-ecosystem. *Ecol. Monogr.* **40**, 23–47 (1970).
- Bouchard, D. C., Williams, M. K. & Surampalli, R. Y. Nitrate contamination of groundwater: sources and potential health effects. *J. Am. Wat. Works Assoc.* **84**, 85–90 (1992).
- Blatt, H. & Jones, R. L. Proportions of exposed igneous, metamorphic, and sedimentary rocks. *Geol. Soc. Am. Bull.* **86**, 1085–1088 (1975).
- Schlesinger, W. H. *Biogeochemistry—An Analysis of Global Change* (Academic, San Diego, California, 1997).
- Goldman, C. R. Primary productivity, nutrients, and transparency during the early onset of eutrophication in ultra-oligotrophic Lake Tahoe, California–Nevada. *Limnol. Oceanogr.* **33**, 1321–1333 (1988).
- Wagner, D. L., Jennings, C. W., Bedrossian, T. L. & Bortugno, E. J. *Geologic Map of the Sacramento Quadrangle, California* (Calif. Div. Geology, Sacramento, California, 1981).
- Clark, L. D. Stratigraphy and structure of part of the western Sierra Nevada metamorphic belt, California. *Prof. Pap. US Geol. Surv.* **410**, (1964).
- Duffield, W. A. & Sharp, R. V. Geology of the Sierra Foothills melange and adjacent areas, Amador County, California. *Prof. Pap. US Geol. Surv.* **827**, (1975).
- Schroeder, P. A. & Ingall, E. D. A method for the determination of nitrogen in clays, with application to the burial diagenesis of shales. *J. Sedim. Res.* **64**, 694–697 (1994).
- Douglas, L. A. in *Minerals in Soil Environments* 2nd edn (eds Dixon, J. B. & Weed, S. B.) 635–674 (Soil Sci. Soc. Am., Madison, Wisconsin, 1989).
- Kulandaiswamy, V. C. & Seetharaman, S. A note on Barnes' method of hydrograph separation. *J. Hydrol.* **9**, 222–229 (1969).
- Caisie, D., Pollock, T. L. & Conjak, R. A. Variation in stream water chemistry and hydrograph separation in a small drainage basin. *J. Hydrol.* **178**, 137–157 (1996).
- Dahlgren, R. A. Comparison of soil solution extraction procedures: effect on solute chemistry. *Commun. Soil Sci. Plant Anal.* **24**, 1783–1794 (1993).
- Bundy, L. G. & Meisinger, J. J. in *Methods of Soil Analysis Part 2, Microbiological and Biochemical Properties* 2nd edn (eds Page, A. L., Miller, R. H. & Kenney, D. R.) 951–984 (Soil Sci. Soc. Am., Madison, Wisconsin, 1982).
- Parfitt, R. L., Percival, H., Dahlgren, R. A. & Hill, L. F. Soil and soil solution chemistry under pasture and radiata pine in New Zealand. *Plant Soil* **191**, 279–290 (1997).
- Aber, J. D., Nadelhoffer, K. J., Steudler, P. & Melillo, J. M. Nitrogen saturation in northern forest ecosystems. *BioScience* **39**, 378–386 (1989).
- Dahlgren, R. A., Singer, M. J. & Huang, X. Oak tree and grazing impacts on soil properties and nutrients in a California oak woodland. *Biogeochemistry* **39**, 45–64 (1997).
- Pavlik, B. M., Muick, P. C., Johnson, S. & Popper, M. *Oaks of California* (Cachuma, Los Olivos, California, 1991).

- Boyd, S. R., Hall, A. & Pillinger, C. T. The measurement of delta ^{15}N in crustal rocks by static vacuum mass spectrometry: application to the origin of the ammonium in the Cornubian batholith, southwest England. *Geochim. Cosmochim. Acta* **57**, 1339–1347 (1993).
- Duit, W., Jansen, J. B. H., van Breemen, A. & Bos, A. Ammonium micas in metamorphic rocks as exemplified by Dome de l'Agout (France). *Am. J. Sci.* **286**, 702–732 (1986).
- Eugster, H. P. & Munoz, J. Ammonium micas: possible sources of atmospheric ammonia and nitrogen. *Science* **151**, 683–686 (1965).
- Hall, A., Pereira, M. D. & Bea, F. The abundance of ammonium in the granites of central Spain, and the behaviour of the ammonium ion during anatexis and fractional crystallization. *Mineral. Petrol.* **56**, 105–123 (1996).
- Kawano, M. & Tomita, K. Mineralogical properties of interstratified ammonium-bearing mica/smectites from Aira, Kagoshima Prefecture, Japan. *Mineral. J.* **15**, 19–31 (1990).
- Krohn, M. D., Kendall, C., Evans, J. R. & Fries, T. L. Relations of ammonium minerals at several hydrothermal systems in the western U.S. *J. Volcanol. Geotherm. Res.* **56**, 401–413 (1993).
- Meyer, F. M. & Ridgway, J. Ammonium in Witwatersrand reefs: a possible indicator of metamorphic fluid flow. *S. Afr. J. Geol.* **1994**, 343–347 (1991).
- Ridgway, J., Appleton, J. D. & Levinson, A. A. Ammonium geochemistry in mineral exploration—comparison of results from the American cordilleras and the southwest Pacific. *Appl. Geochem.* **5**, 475–489 (1990).
- Stevenson, F. J. Chemical state of the nitrogen in rocks. *Geochim. Cosmochim. Acta* **26**, 797–809 (1962).
- SYSTAT v. 6.0 for Windows (SPSS Inc., Chicago, Illinois, 1996).

Acknowledgements. We thank Georgia Pacific for helicopter access to high-elevation watersheds during the winter; East Bay Municipal Utilities District, Georgia Pacific, and the US Forest Service for the loan of streamwater autosamplers; J. Munn, B. McGurk, L. Costock, J. Pierner, B. Smith and the Mokelumne River Water Quality Monitoring Committee for logistical support; and J. Jimenez, D. Levine, R. Northup, G. Pauly, J. Shaw, E. Suzuki and D. Walsh for assistance with field work and analysis. This work was supported by the California Dept of Forestry and Fire Protection.

Correspondence and requests for materials should be addressed to R.A.D. (e-mail: radahlgren@ucdavis.edu).

Cenozoic magmatism throughout east Africa resulting from impact of a single plume

C. J. Ebinger*† & N. H. Sleep*

* Department of Geophysics, Mitchell Building, Stanford University, Stanford, California 94305, USA

† Department of Earth Sciences, University of Leeds, Leeds LS2 9JT, UK

The geology of northern and central Africa is characterized by broad plateaux, narrower swells and volcanism occurring from ~45 Myr ago to the present. The greatest magma volumes occur on the >1,000-km-wide Ethiopian and east African plateaux, which are transected by the Red Sea, Gulf of Aden and east African rift systems, active since the late Oligocene epoch. Evidence for one or more mantle plumes having impinged beneath the plateaux comes from the dynamic compensation inferred from gravity studies, the generally small degrees of extension observed and the geochemistry of voluminous eruptive products^{1–4}. Here we present a model of a single large plume impinging beneath the Ethiopian plateau that takes into account lateral flow and ponding of plume material in pre-existing zones of lithospheric thinning⁵. We show that this single plume can explain the distribution and timing of magmatism and uplift throughout east Africa. The thin lithosphere beneath the Mesozoic–Palaeogene rifts and passive margins of Africa and Arabia guides the lateral flow of plume material west to the Cameroon volcanic line and south to the Comoros Islands. Our results demonstrate the strong control that the lithosphere exerts on the spatial distribution of plume-related melting and magmatism.

The alkali volcanism without rifting on the ~400-km-wide topographic swells and ocean island chains of Africa (Fig. 1) have been attributed to separate plumes^{1,6,7}. Burke¹ suggested that the unusually high elevation and Cenozoic volcanism stems from the African plate's standstill relative to mantle circulation since ~35 Myr ago, which allowed ~40 mantle plumes to penetrate the plate without the creation of hotspot tracks. The Hoggar, Tibesti, Darfur and Adamawa swells and Cenozoic eruptive centres lie near

Mesozoic rifts, and a Mesozoic–Palaeogene rift system underlies the topographic depression between the east African and Ethiopian plateaux^{6,8} (Fig. 1). Wilson and Guiraud⁶ have suggested preferential melting of mantle metasomatized during the Mesozoic breakup of Gondwana, in the presence of many plumes.

Seismic velocity and geochemical anomalies suggest that hot upper mantle underlies the Red Sea, Afar and Eastern rifts^{9–11}. Tomographic models show >200-km-thick lithosphere beneath the cratonic core of east Africa and steep gradients along its margins¹¹, but little is known of upper-mantle structure beneath the Ethiopian plateau away from the rifts (Fig. 1). Marty *et al.*² have reported high ³He/⁴He ratios throughout a 2,000-km-wide region centred on the Ethiopian plateau, indicating a broad mantle thermal anomaly. The largest geoid anomalies coincide with the Ethiopian and east African plateaux where more than 8×10^5 km³ of primarily basaltic material has erupted^{3,12} since ~45 Myr ago. The earliest Cenozoic flood basaltic magmatism occurred in the Ethiopian plateau region, although the Cameroon chain onshore has seen discontinuous activity since Late Cretaceous time^{1,13} (Fig. 1).

Incomplete thermal equilibration of thinned lithosphere beneath Late Cretaceous–Palaeogene rifts and Mesozoic passive margins would have caused significant variations in the depth to the 1,300 °C isotherm (base of lithosphere) beneath Africa 45 Myr ago, particu-

larly in basins active during Palaeogene time (Fig. 1). This topographic relief at the lithosphere–asthenosphere boundary before the arrival of a plume head may have deflected plume material away from its centre, as would cratonic roots. In the numerical models of Sleep^{5,14}, topographic relief at the lithosphere–asthenosphere boundary disrupts the radial spread of buoyant plume material, leading to ponding below thinner lithosphere. Melting is enhanced where plume material cascades over steep relief at the lithosphere–asthenosphere boundary⁵.

The question is, could melt generated from the impact of one strong plume beneath the Ethiopian plateau flow along the elevated lithosphere–asthenosphere boundary beneath the Mesozoic–Palaeogene rifts and passive margins, explaining the swells with small volumes of erupted magma shown in Fig. 1. We use the methods of Sleep⁵ to investigate the effects of pre-existing lithospheric thickness variations on the lateral flow of plume material. Our predictions of the spatial distribution of plume material have implications for the reactivation of pre-plume rifts and the preservation of cratonic roots.

Below and in Figs 1 and 2a we summarize constraints on lithospheric structure before and after initial magmatism in the Ethiopian plateau region. Mantle xenolith and seismic data indicate that Archaean–early Proterozoic lithosphere was >150 km thick

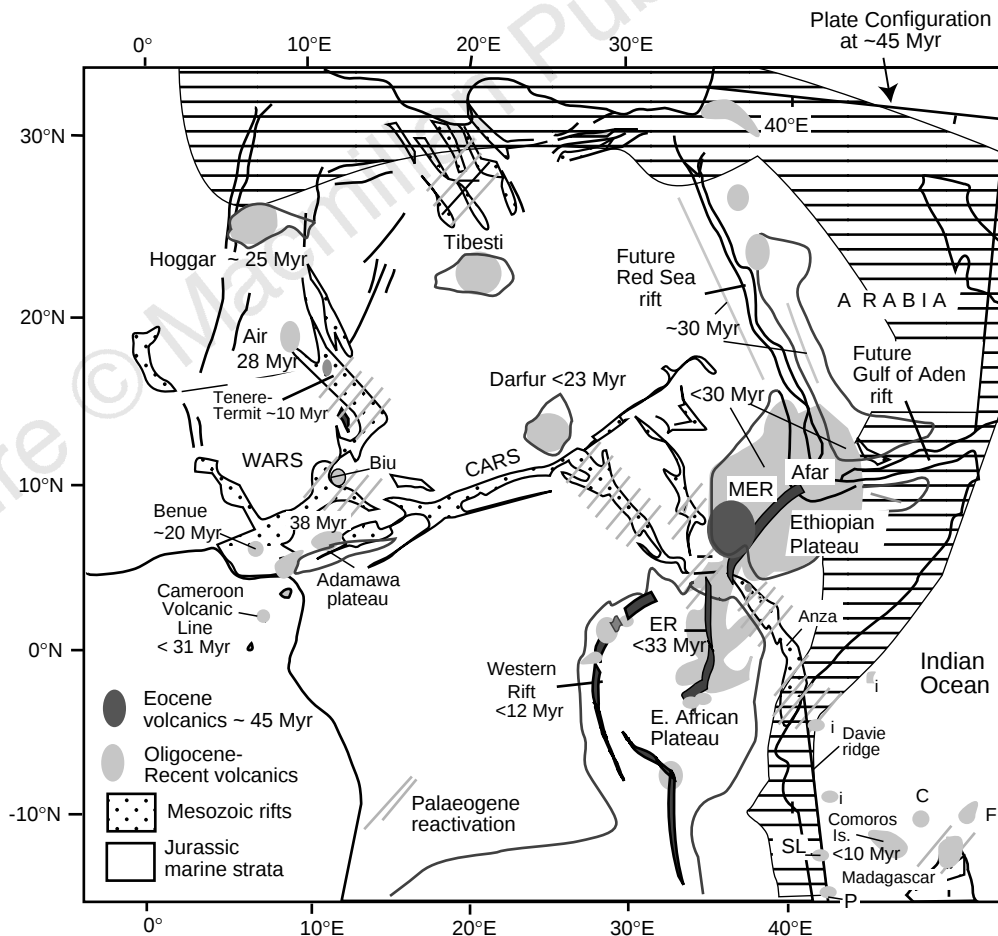


Figure 1 Reconstruction of the African plate at ~45 Myr ago, before the separation of Africa and Arabia. Shown are Cenozoic plateaux, swells, rifts (ER, Eastern rift; MER, Main Ethiopian rift) and magmatism (maximum ages in each province indicated), the distribution of Mesozoic–Palaeogene rifts, and marine strata along passive margins (after ref. 18). Elevations >1,000 m are enclosed by bold lines; uplift (where known) occurred after ~32 Myr ago (ref. 17). Where known from borehole data, peak extension in pre-plume basins occurred in Cretaceous time^{8,16,18,34}. Seismic reflection and borehole data indicate renewed extension during Eocene–recent time in the west African (WARS), central African (CARS)

and Anza rift systems^{8,16,17,31,34}. Lithospheric extension along the Indian Ocean passive margin occurred in Jurassic time; sea-floor spreading between Africa and Madagascar occurred between 165 and 80 Myr ago (ref. 35). Volcanism and extension have also affected the passive margins of east Africa since ~35 Myr ago with reactivation of some Mesozoic structures³⁵ (for example, Davie ridge). Volcanoes of the Comoros chain have developed on oceanic crust since 5 Myr ago, and the Cosmoledo (C), Farquhar (F), St Lazaire (SL) and Paisley (P) seamounts are of similar age^{7,35}; i indicates intrusives.

beneath parts of central and northern Africa 45 Myr ago (refs 11, 15). We assume that cratonic lithosphere was ~ 150 km thick; thicker lithosphere would only enhance the lateral flow effects we describe below. We use extension estimates, where available, to constrain lithospheric thickness beneath the Mesozoic–Palaeogene rifts and passive margins, assuming uniform crust and upper-mantle extension (Fig. 2a). Sequences within many of these basins show a thermal subsidence phase, indicating that the mantle lithosphere was thinned^{18,16–18}.

Volcanism in southwestern Ethiopia 40–45 Myr ago occurred ~ 20 Myr before the earliest known extension in this area^{12,19} (Fig. 1). Between ~ 32 and 25 Myr ago, volcanism was widespread throughout the Red Sea and Gulf of Aden, northern Main Ethiopian rift (MER), and central Ethiopian plateau^{17,19–22} (Fig. 1). The geochemical signature of the plume now extends into the Red Sea and Gulf of Aden^{2,23}, and seismic data suggest crustal underplating beneath the 2,500-m-high eastern plateau⁹.

Far-field stresses generated by the Africa–Eurasia collision may have initiated extension in parts of the Red Sea and Gulf of Aden rifts ~ 30 Myr ago^{17,20}. Sea-floor spreading commenced ~ 20 Myr ago in the Gulf of Aden, propagating westward into Afar by ~ 5 Myr ago (ref. 17). Alkaline volcanism and faulting have propagated southward along the Eastern and Western rifts since 25 Myr ago (Fig. 1). Where timing constraints exist, extension within any province, if present, occurs after initial uplift and volcanism, except in the pre-plume rift zones and along passive margins^{17,19–22}. Overall, we see a sub-radial distribution of magmatism about southwestern Ethiopia since ~ 45 Myr ago.

Hot, low-density material within starting mantle plumes rises diapirically through the mantle and interacts with the base of the lithosphere, forming a broad head that may partially melt to produce flood basalts and radial dykes^{24,25}. This buoyant material displaces denser, normal asthenosphere and heats the lithosphere,

leading to regional uplift, lithospheric thinning and pressure-release magmatism. General models differ in the interactions between plume and lithosphere and the timing of pressure-release volcanism, with variations in plume temperatures of ~ 10 K leading to significant variations in melt volumes^{26,27}. Sleep^{5,14} shows that pre-existing zones of lithospheric thinning act as sinks for buoyant plume material, driving lateral flow hundreds of kilometres from the plume's centre.

Using the analytical and numerical methods of Sleep^{5,14}, we delineate areas of ponded plume material that are prone to pressure-release magmatism. Assuming a linear depth versus melting parametrization²⁸, we can predict final melt thickness above the plume. We do not distinguish between plume head and tail material, given the slow movement of the African plate since 45 Myr ago. We assume that the plume initially was centred at 38° E, 8° N where the earliest basalts are found, and that it moves to 36° E, 4° N over 45 Myr. Very similar results are obtained if the head is moved a few degrees in any direction.

The numerical model predicts the distribution of plume material and melt beneath central and northern Africa after onset (Fig. 2b, c). At ~ 45 Myr ago a circular plume head of 750 km radius impinges on the base of the lithosphere, producing a thick, low-density layer beneath the Ethiopian plateau. There is a net upward movement of laterally flowing plume material, with steep gradients increasing vertical velocity and pressure-release melting^{5,26} (Fig. 2c). Considering the thick lithospheric lid, local factors (for example, pre-existing weaknesses and volatile content) probably control eruptive centre locations.

By 40 Myr ago, plume material has moved an additional 500–800 km southward and eastward to thinned lithosphere along the Indian Ocean margin and beneath the Arabian peninsula, following Mesozoic–Palaeogene rifts. Between 35 and 10 Myr ago a narrow tongue of plume material ‘trickles’ westward beneath the central

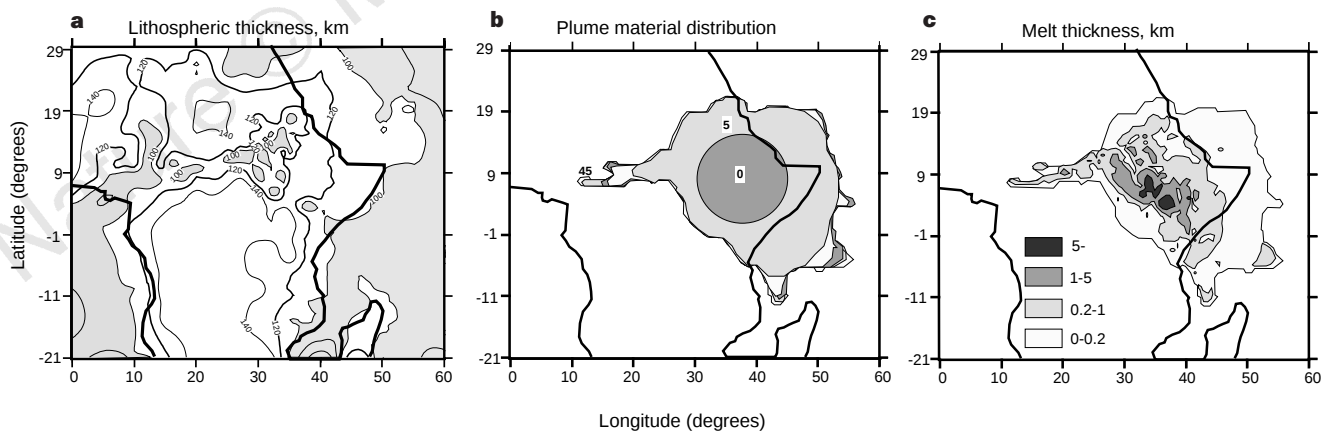


Figure 2 Results of plume model. **a**, Model of pre-plume lithospheric thickness (values <100 km are shaded) at ~ 45 Myr ago constrained by estimates of crustal stretching, seismic^{9–11,32} and xenolith^{15,30} data. Maximum crustal extension in Mesozoic–Palaeogene rift basins of the west and central African rift systems varies from $<10\%$ to $\sim 50\%$, based on subsidence patterns, fault reconstructions, and models of gravity data^{8,16–18,34,35}. The northern and northeastern margins of Africa and much of the Arabian peninsula were passive margins into Palaeogene time; we assume 110–100-km-thick lithosphere here. Africa's northeastern margin was affected by folding in Late Cretaceous and Eocene time with the collision of Africa and Eurasia¹; we assume a ~ 115 -km-thick lithosphere here. Oceanic lithosphere along Africa's eastern and western margins is assumed to follow a normal thermal subsidence curve until the onset of the plume at ~ 45 Myr ago, with formation ages assigned from the model of Coffin and Rabinowitz²⁵. **b**, Contours of the maximum horizontal extent of plume material between 45 Myr ago (contour 0) and present (contour 45) predicted by the interactions of plume

with the lithospheric lid shown in **a**. We note the lobe of plume material which roughly coincides with the Indian Ocean seamounts and intrusives, and the tongue of material beneath the central African rift system (Fig. 1). We use the same model parameters as Sleep¹⁴, except that we assume a starting plume head volume of $0.2 \times 10^{18} \text{ m}^3$, a viscosity of $0.3 \times 10^{18} \text{ Pa s}$ and a constant buoyancy flux of 4 Mg s^{-1} similar to that of a strong plume²⁷. Only local buoyancy is considered. As shown in Sleep¹⁴, the depth to the base or viscosity of the ordinary asthenospheric channel and its replenishment mechanism have little effect on results. This model reproduces the spreading drop similarity solution³⁶ to 0.5% at 45 Myr. **c**, The equivalent thickness of melt produced from 30 Myr ago to the present. The calculation includes pressure-release melting from lateral flow of material and lithospheric thinning caused by the interaction of plume material and the base of the lithosphere. Plume material is assumed to melt by 0.1% per kilometre of ascent, independent of depth. Melting is enhanced above steep gradients shown in **a**.

African rift zone, reaching the western Adamawa plateau by ~5 Myr ago. Our predictions for this region are probably minima due to the earlier history of plume activity and rifting in this area.

This simple model of lithosphere–plume interactions in Africa correlates well with the distribution and timing of magmatism in east and central Africa, and along the Indian Ocean margin (Figs 1, 2). For example, ponding and melting occur beneath the Darfur swell by 30 Myr ago, and beneath the Davie ridge by 5 Myr ago, explaining the young seamounts along this Mesozoic structure (Figs 1, 2). The radial flow of plume material away from the plume's centre fits well with the observed southward propagation of volcanism in east Africa since Late Oligocene time (Fig. 1). The Ethiopian plume model cannot explain the Tibesti or Hoggar swells, unless undiscovered rifts lie beneath the Sahara. We could increase the plume volume, lower viscosity, and modify initial lithospheric thickness to increase the radial flow, but too little information on present and ancient lithospheric structure exists to justify refinements of our simple model.

Our model includes only thermal thinning of the lithosphere above the plume material, but we can readily anticipate the model's response to lithospheric stretching within the Red Sea and Gulf of Aden. Newly thinned lithosphere would have channelled melt from the eastern limbs into the axis of the Red Sea and Aden rifts, as well as adjacent margins. This extension along the Arabian margins at ~30 Myr ago would have triggered decompression melting of ponded plume material, causing widespread magmatism along the Arabian margins. The superposition of Red Sea, Aden and east African extension in Afar may have caused greater pressure-release melting there rather than above our plume's centre.

In active rifting models, lateral density variations within the upper mantle generate extensional stresses large enough to initiate rifting²⁹. Adiabatic decompression melting beneath the rift enhances flow of plume material, and rifts propagate. This mechanism may explain the extension after magmatism and uplift, and the southward propagation of rifting in east Africa. Although it is an indirect result, we suggest that episodic extension in the Red Sea and Gulf of Aden captured ponded melt, and led to pulses of basaltic magmatism in east Africa. Reactivation of pre-plume basins and initial extension in the northern Kenya rift ~25 Myr ago would have created a conduit for the distribution of plume material southward beneath the Eastern and Western rift systems (Figs 1, 2a). This additional extension not included in the model would increase outward flow, perhaps as far south as Madagascar and the Western rift where magmatic centres are located near Mesozoic rifts (Fig. 1).

The steep gradients at the lithosphere–asthenosphere boundary along the Tanzania craton/mobile-belt boundaries¹¹ may focus flow, explaining <30-Myr-old carbonatitic and kimberlitic magmas derived from small melt volumes at depths ≥150 km (refs 15, 30). The pre-plume relief at the lithosphere–asthenosphere boundary beneath the east African cratons diverts plume material and focuses decompression melting along the craton margins, preserving their thick, strong cores (Fig. 2c).

These results have implications for the rejuvenation and thermal subsidence of continental rifts. Ponded plume material contributes to regional uplift, but sills intruded into sedimentary strata or magma underplated at the base of previously thinned crust can locally augment or counter this uplift, depending on density contrasts. Many Mesozoic–Palaeogene rift basins lying along the predicted paths of plume material show Late Cenozoic reactivation and igneous intrusions^{16,31}, but borehole and seismic data are available from many basins. Coffin *et al.*³² report anomalous upper-mantle velocities of 7.8 km s⁻¹ beneath the western Indian Ocean basin, suggesting that mantle temperatures are elevated above the southeastern lobe (Fig. 2c).

Melt composition will vary depending on plume interactions with the heated lithosphere and normal asthenosphere, which will increase with distance from the source. Temporal variations will

occur due to compositional and temperature differences between plume head and tail³³.

Our studies demonstrate that the location of continental flood basalts may not coincide with the centre of the starting plume head, and several discrete magmatic provinces can be produced by a single plume. This numerical model of the Ethiopian plume also supports a plume origin for the east African rift system. □

Received 28 July 1997; accepted 6 August 1998.

- Burke, K. The African plate. *S. Afr. J. Geol.* **99**, 339–410 (1996).
- Marty, B., Pik, R. & Gezahagen, Y. He isotopic variations in Ethiopian plume lavas: nature of magmatic sources and limit on lower mantle convection. *Earth Planet. Sci. Lett.* **144**, 223–237 (1996).
- Latin, D., Norry, M. & Tarzey, R. Magmatism in the Gregory rift: evidence for melt generation by a plume. *J. Petrol.* **34**, 1007–1027 (1993).
- Ebinger, C., Bechtel, T., Forsyth, D. & Bowin, C. Effective elastic plate thickness beneath the East African and Afar plateaux, and isostatic compensation for the uplifts. *J. Geophys. Res.* **94**, 2893–2901 (1989).
- Sleep, N. H. Lateral flow of hot plume material ponded at sub-lithospheric depths. *J. Geophys. Res.* **101**, 28065–28083 (1996).
- Wilson, M. & Giraud, R. Magmatism and rifting in western and central Africa, from Late Jurassic to recent times. *Tectonophysics* **213**, 203–225 (1993).
- Emerick, C. M. & Duncan, R. Age progressive volcanism in the Comores Archipelago, western Indian Ocean and implications for Somali plate motion. *Earth Planet. Sci. Lett.* **60**, 415–428 (1982).
- Hendrie, D., Kusznir, N., Morley, C. K. & Ebinger, C. A quantitative model of rift basin development in the northern Kenya rift: evidence for the Turkana region as an "accommodation zone" during the Palaeogene. *Tectonophysics* **236**, 409–438 (1994).
- Prodehl, C. & Mechie, J. Crustal thinning in relationship to the evolution of the Afro-Arabian rift system: A review of seismic refraction data. *Tectonophysics* **198**, 311–327 (1991).
- Green, V., Achauer, U. & Meyer, R. A three-dimensional image of the crust and upper mantle beneath the Kenya rift. *Nature* **354**, 199–203 (1991).
- Ritsema, J., Nyblade, A., Owens, T., Langston, C. & VanDecar, J. Upper mantle velocity structure beneath Tanzania, E. Africa: implications for the stability of cratonic lithosphere. *J. Geophys. Res.* (in the press).
- George, R. *Thermal and Tectonic Controls on Magmatism in the Ethiopian Rift*. Thesis, Open Univ. (1997).
- Lee, D. C., Halliday, K. A. N., Fitton, J. G. & Pol, G. Chronology of volcanism along the Cameroon volcanic line. *Earth Planet. Sci. Lett.* **123**, 119–138 (1994).
- Sleep, N. Lateral flow and ponding of starting plume material. *J. Geophys. Res.* **102**, 10001–10012 (1997).
- Boyd, F. & Gurney, J. Diamonds in the African lithosphere. *Science* **232**, 272–477 (1986).
- Genik, G. Regional framework, structure, and petroleum aspects of rift basins in Niger, Chad, and Central African Republic. *Tectonophysics* **213**, 169–185 (1992).
- Menzies, M., Gallagher, K., Yelland, A. & Hurford, A. Volcanic and nonvolcanic rifted margins of the Red Sea and Gulf of Aden: crustal cooling and margin evolution in Yemen. *Geochim. Cosmochim. Acta* **61**, 2511–2527 (1997).
- Bosworth, W. Mesozoic and Tertiary rifting in East Africa. *Tectonophysics* **209**, 115–137 (1992).
- Ebinger, C. J., Yemane, T., WoldeGabriel, G. & Aronson, J. Late Eocene–Recent volcanism and faulting in the southern Main Ethiopian rift system. *J. Geol. Soc. Lond.* **150**, 99–108 (1993).
- Baker, J., Snee, L. & Menzies, M. A brief Oligocene period of flood volcanism in Yemen: implications for the duration and rate of continental flood volcanism at the Afro-Arabian triple junction. *Earth Planet. Sci. Lett.* **138**, 39–55 (1996).
- Morley, C. K. *et al.* Tectonic evolution of the northern Kenya rift. *J. Geol. Soc. Lond.* **149**, 333–348 (1992).
- Hofmann, C. Timing of the Ethiopian flood basalt event and implications for plume birth and global change. *Nature* **389**, 838–841 (1997).
- Schilling, J.-G., Kingsley, R., Hanan, B. & McCully, B. Nd–Sr–Pb isotopic variations along the Gulf of Aden: evidence for the Afar mantle plume–lithosphere interaction. *J. Geophys. Res.* **97**, 10927–10966 (1992).
- Griffiths, R. & Campbell, I. Interaction of mantle plume heads with Earth's surface at the onset of small-scale convection. *J. Geophys. Res.* **96**, 18295–18310 (1991).
- Hill, R. I. Starting plumes and continental break-up. *Earth Planet. Sci. Lett.* **104**, 398–416 (1991).
- White, R. & McKenzie, D. Magmatism at rift zones: the generation of volcanic continental margins and flood basalts. *J. Geophys. Res.* **94**, 7685–7729 (1989).
- Ribe, R. & Christensen, U. Three-dimensional modeling of plume–lithosphere interaction. *J. Geophys. Res.* **99**, 669–682 (1994).
- Watson, S. & McKenzie, D. Melt generation by plumes: a study of Hawaiian volcanism. *J. Petrol.* **32**, 501–537 (1991).
- Crough, S. T. Thermal origin of midplate swells. *Geophys. J. R. Astron. Soc.* **55**, 451–469 (1978).
- Dawson, B. & Smith, M. Veined and metasomatised peridotites from Pello Hill, Tanzania: evidence for anomalously light mantle beneath the Tanzania sector of the eastern rift valley. *Contrib. Mineral. Petrol.* **100**, 510–527 (1988).
- Bosworth, W. & Morley, C. K. Structural and stratigraphic evolution of the Anza rift, Kenya. *Tectonophysics* **236**, 93–115 (1994).
- Coffin, M. F., Rabinowitz, P. & Houtz, R. E. Crustal structure in the western Somali basin. *Geophys. J. R. Astron. Soc.* **86**, 331–365 (1986).
- Richards, M., Duncan, R. & Courtillot, V. Flood basalts and hotspot tracks: plume heads and tails. *Science* **246**, 103–107 (1989).
- van der Meer, F. & Cloetingh, S. Intraplate stresses and the subsidence history of the Sirte basin (Libya). *Tectonophysics* **226**, 37–58 (1993).
- Coffin, M. F. & Rabinowitz, P. *Evolution of the Conjugate Madagascar and W. Somali Basin* (Spec. Publ. 226, Geol. Soc. Am., Boulder, Colorado, (1988).
- Huppert, H. E. The propagation of two-dimensional axisymmetric viscous gravity currents over a rigid horizontal surface. *J. Fluid Mech.* **121**, 43–58 (1982).

Acknowledgements. C.J.E. completed this work while on sabbatical leave at Stanford University. We thank G. Davies for detailed comments, and M. Menzies, G. Thompson, K. Burke, C. Ruppel, T. Parsons, R. George, G. Barth and J. Cann for discussions, which improved this work.

Correspondence and requests for materials should be addressed to C.J.E. at the University of Leeds (e-mail: cindy@earth.leeds.ac.uk).

Osteolepiforms and the ancestry of tetrapods

Per E. Ahlberg* & Zerina Johanson†

* Department of Palaeontology, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

† Palaeontology Section, Australian Museum, 6 College Street, Sydney South, New South Wales 2000, Australia

Fossil discoveries^{1–7} and improved phylogenies^{3–5,7} have greatly improved our understanding of the origin of tetrapods, making it possible to reconstruct sequences of character change leading to tetrapod morphologies^{5,7} and to tentatively identify the genetic basis for some of these changes^{8,9}. However, progress has centred on the upper part of the Tetrapodomorpha⁵ which is occupied by Devonian tetrapods such as *Acanthostega*^{1,2,5} and *Ichthyostega*¹. Few advances have been made in improving our understanding of the lower, 'fish' part of the group, beyond establishing Elpistostegalia, Osteolepiformes and Rhizodontida as progressively more

primitive constituents^{10–13}. It has not been convincingly confirmed or disproved that the Osteolepiformes, a diverse but structurally uniform group that is central to the debate about tetrapod origins^{14–17}, is monophyletic relative to tetrapods (that is, a single side branch on the tetrapod lineage). The earliest steps of the fish–tetrapod transition have thus remained poorly resolved. Here we present the first detailed analysis of the lower part of the Tetrapodomorpha, based on 99 characters scored for 29 taxa. We show that both the Osteolepiformes as a whole and their constituent group Osteolepididae are paraphyletic to tetrapods (that is, each comprises a section of the tetrapod lineage with several side branches), and that their 'uniting characters' are attributes of the tetrapodomorph stem lineage. The supposedly discredited idea of osteolepiforms as tetrapod ancestors^{14–17} is, in effect, supported by our analysis. Tetrapod-like character complexes evolved three times in parallel within the Tetrapodomorpha.

During much of this century, research into the origin of Tetrapoda (defined here provisionally as 'vertebrates with limbs', thus including the crown group and top end of the stem of the total group Tetrapodomorpha^{3,12}) focused on the osteolepiforms as potential ancestors^{14–17}. Particular emphasis was placed on *Eusthenopteron*, the first osteolepiform to be described in exhaustive anatomical detail^{14,16}. *Eusthenopteron* was frequently used as a starting point for explaining the origin of tetrapod characters^{14,16,17}. However, the first thorough cladistic review of sarcopterygian interrelationships¹⁸ dismissed the osteolepiforms as an ill-defined assemblage of primitive lobe-fins, remote from tetrapods. Further work reinstated the osteolepiforms in the tetrapod stem group^{10,19,20}, but has generally avoided the issue of osteolepiform

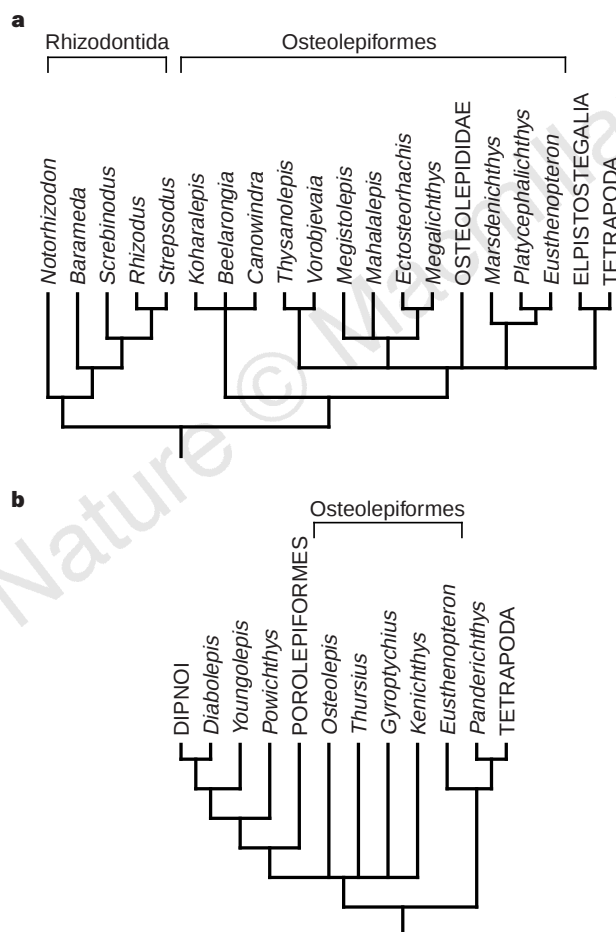


Figure 1 Previous osteolepiform phylogenies. **a**, Young *et al.*'s analysis²¹. **b**, Chang and Yu's analysis²². The elpistostegid–tetrapod sistergroup relationship is the only point of agreement. Both cladograms contain large unresolved polychotomies. Young *et al.*'s analysis (produced 'by hand') is based on 38 characters, of which 20 match characters in our matrix, 9 are inapplicable to our analysis, and 9 were rejected by us. Chang and Yu's analysis (produced by PAUP3.1.1) is based on 90 characters, of which 48 match ours, 21 are inapplicable, and 21 were rejected. Judgements of character states for certain taxa differ from ours in both cases.

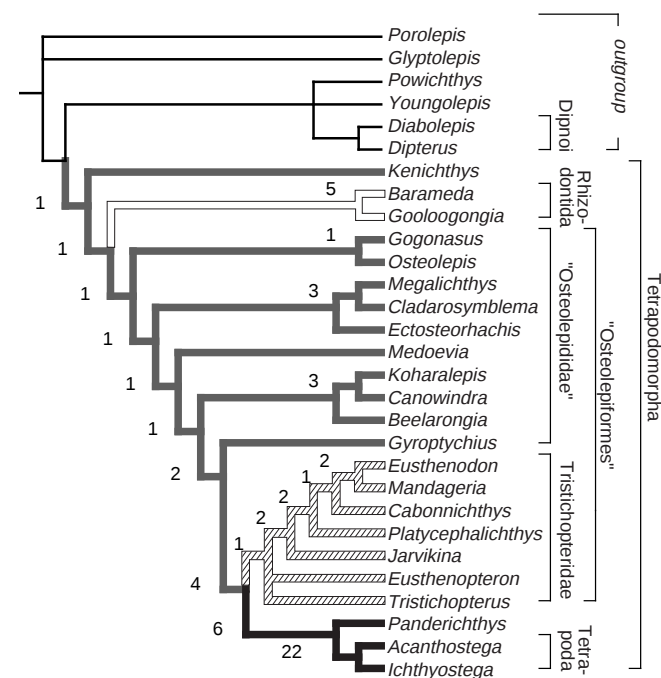


Figure 2 New phylogeny of the tetrapodomorph stem group (consensus of 9 trees, based on 99 characters scored for 29 taxa). Bremer support values are shown at nodes. Tree length, 196 steps; consistency index, 0.571; homoplasy index, 0.429; retention index, 0.772; rescaled consistency index, 0.441. Thin black lines represent outgroups; grey lines indicate 'osteolepids' (including the problematic genus *Kenichthys*³⁰); empty outline indicates rhizodonts; positive slope hatching indicates tristichopterids; thick black lines indicate elpistostegids+tetrapods. 'Tetrapoda' indicates the clade 'vertebrates with limbs' which contains, but is more inclusive than, the tetrapodomorph crown group.

monophyly by using *Eusthenopteron*, explicitly or implicitly, as the principal representative.

Two analyses have attempted to resolve osteolepiform interrelationships cladistically, with incompatible results (Fig. 1)^{21,22}. Both were based on relatively small data sets of 760 and 1,260 data points, respectively, as compared with 2,871 for our analysis. More important, one analysis²¹ included the 'rhizodont' *Notorhizodon*, which appears to be a mixture of at least two unrelated fishes (P.E.A. and Z.J., manuscript in preparation), whereas the other²² included published data on *Thursius*²³, which contain some inaccuracies because of preparation damage (P.E.A., personal observation). Our analysis is also the first to incorporate the perfectly preserved three-dimensional osteolepiforms *Medoevia*²⁴ and *Gogonassus*²⁵, and the articulated rhizodont *Gooloogongia*¹³, which add greatly to the quality of the data.

Our analysis (Fig. 2) shows the Osteolepiformes to be paraphyletic. Of the two traditionally accepted osteolepiform subgroups, the Tristichopteridae (=Eustenopteridae) are closer to tetrapods than are the Osteolepididae, and the latter are themselves paraphyletic. Bremer support values for the nodes are generally low (Fig. 2). However, specifying a monophyletic Osteolepiformes increases the tree length by six steps. Reversing the positions of osteolepids and tristichopterids relative to tetrapods, by specifying an osteolepid–elpistostegid–tetrapod clade, likewise requires six steps. In contrast, a monophyletic Osteolepididae can be achieved with only two extra steps. We conclude that osteolepiform paraphyly relative to Tetra-

poda+Elpistostegalia, and the position of Tristichopteridae crownward to Osteolepididae, is well supported, but that the support for osteolepidid paraphyly is weak. Osteolepiformes should no longer be used as a formal taxonomic group.

The implications of this are profound. If the 'osteolepiforms' are paraphyletic, their numerous shared characteristics are attributes of the stem lineage, and thus 'ancestral characters' for the Tetrapoda+Elpistostegalia. The characters include *inter alia* (Fig. 3) a hinged braincase where the hinge ran through the profundus nerve foramen, a unique pattern of dermal bones, a small tripodal scapulocoracoid (endoskeletal shoulder girdle), anal and posterior dorsal fin supports comprising a basal plate and three unjointed radials, and a pectoral fin skeleton comprising four axial elements (humerus, ulna, ulnare, IV), preaxial unjointed radials and a postaxial flange on the ulnare^{14–17}.

This information allows more precise statements to be made about the evolution of tetrapod morphologies, and will thus allow more rigorous framing of hypotheses in tetrapod phylogenetics, biomechanics and evolutionary developmental genetics. The paraphyletic condition of the 'Osteolepiformes' relative to Tetrapoda and the comparatively crownward position of *Eusthenopteron* both agree surprisingly well with precladistic perceptions of this group^{14–16}.

Derived characters which originate within the 'osteolepiform' part of the Tetrapodomorpha include loss of cosmine, loss of extratemporal bones, narrowing of the otic part of the skull, and lengthening of the snout, orbitotemporal region and corresponding parts of the lower jaw. Most of these characters appear at the ((Tetrapoda+Elpistostegalia) Tristichopteridae) node (Fig. 2). The last character complex may be functionally associated with the reduction in intracranial joint mobility and the adoption of a snapping mode of prey capture.

The 'osteolepiform' part of the stem lineage probably consisted

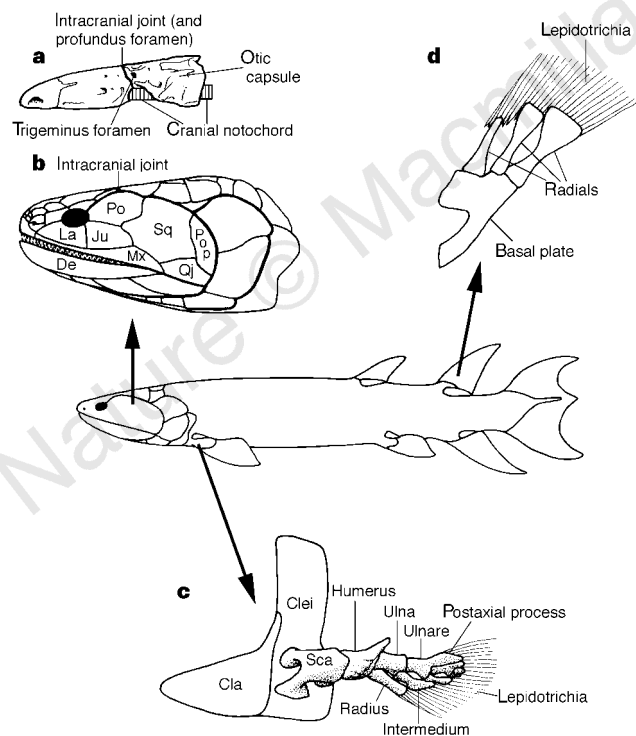


Figure 3 Ancestral characters for clade Elpistostegalia+Tetrapoda. These precise morphologies, illustrated by *Eusthenopteron* (modified from ref. 16), characterize the tetrapod stem lineage between the *Osteolepis*+*Gogonassus* node and the Tristichopteridae node. **a**, Braincase (known in many genera, including *Osteolepis*, *Gogonassus*, *Ectosteorachis* and *Medoevia*). **b**, Dermal skull bones (known in most genera). La, lacrima; Ju, jugal; Po, postorbital; Sq, squamosal; Q, quadrate; Pop, preopercular; De, dentary; Mx, maxilla. **c**, Pectoral girdle and fin, mesial view (pectoral fin skeleton known in *Megalichthys*, *Sterropterygion* and tristichopterids; girdle known in many genera). Cla, clavicle; clei, cleithrum; Sca, scapulocoracoid. **d**, Posterior dorsal fin support (known in *Megalichthys*, *Rhizodopsis* and tristichopterids). (Redrawn by permission of the Royal Society of Edinburgh from ref. 14.)

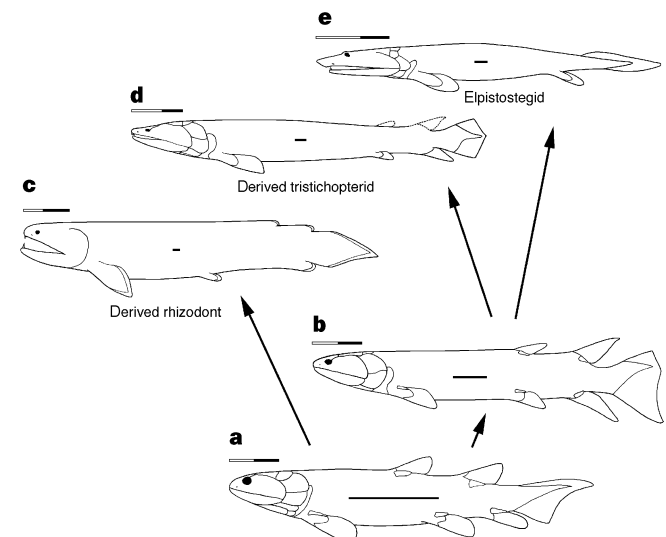


Figure 4 Parallel evolution in the tetrapod stem group, inferred from the phylogeny in Fig. 2. **a**, The 'osteolepidid' *Osteolepis* (modified from ref. 16), approximating to the stem-lineage condition between the *Kenichthys* and *Gyropterychius* nodes. **b**, The tristichopterid *Tristichopterus* (from ref. 27) approximating to the common ancestor of the clade (Tristichopteridae (Elpistostegalia+Tetrapoda)). **c**, The rhizodont *Strepsodus* (modified from ref. 26). **d**, The tristichopterid *Mandageria* (from ref. 27). **e**, The elpistostegid *Panderichthys* (modified from ref. 28). Scale bars within fishes, 5 cm; bars above heads show the length of the postparietal bones (black) as a proportion of the total head length. The intracranial joint lies at the anterior margin of postparietals. Rhizodont skull proportions are based on *Barameda* (ref. 11). (Redrawn by permission of the Royal Society of Edinburgh from ref. 29.)

of rather generalized, small (20–70 cm in length) fishes, with dentitions lacking anterior fangs, a morphological category represented in the analysis by *Kenichthys*, *Gogonassus*, *Osteolepis* (Fig. 4a), *Medoevia*, the canowindrids and *Tristichopterus* (Fig. 4b). Rhizodonts²⁶, derived tristichopterids²⁷ and elpistostegids+tetrapods^{5,16,28}, in contrast, show parallel trends towards a quite different morphology: they increased dramatically in size, reduced or lost their median fins, acquired diphyrcal tails with a low aspect ratio, and developed a pair of fangs at the lower jaw symphysis (Fig. 4c–e). Rhizodonts and derived tristichopterids also acquired premaxillary fangs^{27,29}. Rhizodonts seem to have retained a primitive, short-snouted skull morphology (J. Jeffery, personal communication). However, tristichopterids and elpistostegids+tetrapods, having a moderately lengthened snout as a synapomorphy (Fig. 4b), independently developed this character further in parallel (Fig. 4d, e). Derived tristichopterids such as *Mandageria*²⁹ (Fig. 4d) have very elpistostegid-like head proportions.

These changes seem to have occurred during the Middle/Late Devonian period in all three groups. Elpistostegids originated in the latest Givetian⁶; the earliest known derived tristichopterid is the Frasnian *Platycephalichthys*^{23,27}; and the earliest known large rhizodont is the Famennian *Sauripteris*¹⁵.

Our analysis indicates that much of the lower part of the tetrapodomorph stem lineage consisted of 'osteolepiform' fishes. The character attributes of this part of the stem lineage can be reconstructed with precision. Parallel evolution towards the morphology of a large predator, with reduced median fins and elaborate anterior dentition, occurred at about the same time in rhizodonts, tristichopterids, and elpistostegids+tetrapods (Fig. 4). The evolution of two latter clades, having extra synapomorphies, also paralleled each other more closely. The Tetrapoda thus arose out of one of several similar evolutionary 'experiments' with a large aquatic predator role. Closer study of these parallel radiations should cast much new light on the ecological background to the origin of tetrapods. □

Methods

Phylogenetic analysis. The analysis was performed using the software package PAUP3.1 with a data matrix of 29 taxa scored for 99 morphological characters (Supplementary information). Characters were scored from specimens or good photographs, not reconstruction drawings. Most parsimonious trees were identified using the heuristic search algorithm, stepwise addition, with 500 random iterations. All characters were weighted equally. Characters 15, 20, 23, 25, 32, 70 and 80 were ordered.

Received 3 June; accepted 20 July 1998.

- Coates, M. I. & Clack, J. A. Polydactyly in the earliest known tetrapod limbs. *Nature* **347**, 66–69 (1990).
- Clack, J. A. Earliest known tetrapod braincase and the evolution of the stapes and fenestra ovalis. *Nature* **369**, 392–394 (1994).
- Ahlberg, P. E. *Elginerpeton pancheni* and the earliest tetrapod clade. *Nature* **373**, 420–425 (1995).
- Lebedev, O. A. & Coates, M. I. The postcranial skeleton of the Devonian tetrapod *Tulerpeton curtum* Lebedev. *Zool. J. Linn. Soc.* **114**, 307–348 (1995).
- Coates, M. I. The Devonian tetrapod *Acanthostega gunnari* Jarvik: postcranial anatomy, basal tetrapod interrelationships and patterns of skeletal evolution. *Trans. R. Soc. Edinb. Earth Sci.* **87**, 363–421 (1996).
- Ahlberg, P. E., Clack, J. A. & Lukševičs, E. Rapid braincase evolution between *Panderichthys* and the earliest tetrapods. *Nature* **381**, 61–64 (1996).
- Ahlberg, P. E. Postcranial stem tetrapod remains from the Devonian of Scat Craig, Morayshire, Scotland. *Zool. J. Linn. Soc.* **122**, 99–141 (1998).
- Shubin, N. The evolution of paired fins and the origin of tetrapod limbs. *Evol. Biol.* **28**, 39–85 (1995).
- Shubin, N., Tabin, C. & Carroll, S. Fossils, genes and the evolution of animal limbs. *Nature* **388**, 638–648 (1997).
- Schultze, H.-P. Dipnoans as sarcopterygians. *J. Morphol.* **1** (Suppl.), 39–74 (1986).
- Long, J. A. A new rhizodontiform fish from the Early Carboniferous of Victoria, Australia, with remarks on the phylogenetic position of the group. *J. Vert. Palaeontol.* **9**, 1–17 (1989).
- Cloutier, R. & Ahlberg, P. E. in *Interrelationships of Fishes* (eds Stiassny, M. L. J., Parenti, L. R. & Johnson, G. D.) 445–479 (Academic, San Diego, 1996).
- Johanson, Z. & Ahlberg, P. E. A complete primitive rhizodont from Australia. *Nature* **394**, 569–572 (1998).
- Andrews, S. M. & Westoll, T. S. The postcranial skeleton of *Eusthenopteron foordi* Whiteaves. *Trans. R. Soc. Edinb.* **68**, 207–329 (1970).
- Andrews, S. M. & Westoll, T. S. The postcranial skeleton of rhipidistian fishes excluding *Eusthenopteron*. *Trans. R. Soc. Edinb.* **68**, 391–489 (1970).
- Jarvik, E. *Basic Structure and Evolution of Vertebrates* Vol. 1 (Academic, New York, 1980).

- Rackoff, J. S. in *The Terrestrial Environment and the Origin of Land Vertebrates* (ed. Panchen, A. L.) 255–292 (Academic, London, 1980).
- Rosen, D. E., Forey, P. L., Gardiner, B. G. & Patterson, C. Lungfishes, tetrapods, palaeontology, and plesiomorphy. *Bull. Am. Mus. Nat. Hist.* **167**, 159–276 (1981).
- Holmes, E. B. Are lungfishes the sister group of tetrapods? *Biol. J. Linn. Soc.* **25**, 379–397 (1985).
- Panchen, A. L. & Smithson, T. S. Character diagnosis, fossils and the origin of tetrapods. *Biol. Rev. Cambridge Phil. Soc.* **62**, 341–438 (1987).
- Young, G. C., Long, J. A. & Ritchie, A. Crossospterygian fishes from the Devonian of Antarctica: systematics, relationships and biogeographic significance. *Rec. Austr. Mus.* **14** (Suppl.), 1–77 (1992).
- Chang, M.-M. & Yu, X. Reexamination of the relationship of Middle Devonian osteolepids—fossil characters and their interpretations. *Am. Mus. Novit.* **3189**, 1–20 (1997).
- Vorobyeva, E. I. 1977. Morphology and nature of evolution of crossospterygian fishes. *Trudy Paleontol. Inst.* **94**, 1–239 (1977).
- Lebedev, O. A. Morphology of a new osteolepidid fish from Russia. *Bull. Mus. Nat. Hist. Nat. C* **17**, 287–341 (1995).
- Long, J. A., Barwick, R. E. & Campbell, K. S. W. Osteology and functional morphology of the osteolepiform fish *Gogonassus andrewsae* Long, 1985, from the Upper Devonian Gogo Formation, Western Australia. *Rec. West. Aust. Mus.* **53** (Suppl.), 1–89 (1997).
- Andrews, S. M. Rhizodont crossospterygian fish from the Dinantian of Foulden, Berwickshire, Scotland, with a re-evaluation of this group. *Trans. R. Soc. Edinb. Earth Sci.* **76**, 67–95 (1985).
- Ahlberg, P. E. & Johanson, Z. Second tristichopterid (Sarcopterygii, Osteolepiformes) from the Upper Devonian of Canowindra, New South Wales, Australia, and phylogeny of the Tristichopteridae. *J. Vert. Palaeontol.* **17**, 653–673 (1997).
- Vorobyeva, E. I. & Schultze, H.-P. in *Origins of the Higher Groups of Tetrapods: Controversy and Consensus* (eds Schultze, H.-P. & Trueb, L.) 68–109 (Cornell Univ., Ithaca, 1991).
- Johanson, Z. & Ahlberg, P. E. New tristichopterid (Osteolepiformes: Sarcopterygii) from the Mandagery Sandstone (Famennian) near Canowindra, N. S. W., Australia. *Trans. R. Soc. Edinb. Earth Sci.* **88**, 39–68 (1997).
- Chang, M.-M. & Min, Z. A new Middle Devonian osteolepid from Qujing, Yunnan. *Mem. Ass. Australasian Palaeontol.* **15**, 183–198 (1993).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. We thank the Australian Museum for their award of a Visiting Fellowship to P.E.A., E. Mark-Kurik for access to material of *Thursius estonicus*, J. Jeffery for information about rhizodonts and Academic Press for Figs 3a–c, 4a.

Correspondence and requests for materials should be addressed to P.E.A. (e-mail: P.Ahlberg@nhm.ac.uk).

Weak trophic interactions and the balance of nature

Kevin McCann, Alan Hastings & Gary R. Huxel

Department of Environmental Sciences and Policy, University of California, Davis, California 95616, USA

Ecological models show that complexity usually destabilizes food webs^{1,2}, predicting that food webs should not amass the large numbers of interacting species that are in fact found in nature^{3–5}. Here, using nonlinear models, we study the influence of interaction strength (likelihood of consumption of one species by another) on food-web dynamics away from equilibrium. Consistent with previous suggestions^{1,6}, our results show that weak to intermediate strength links are important in promoting community persistence and stability. Weak links act to dampen oscillations between consumers and resources. This tends to maintain population densities further away from zero, decreasing the statistical chance that a population will become extinct (lower population densities are more prone to such chances). Data on interaction strengths in natural food webs^{7–11} indicate that food-web interaction strengths are indeed characterized by many weak interactions and a few strong interactions.

Here we combine formally the influence of interaction strength with modern food-web data and models, uniting verbal arguments^{12–16} with the rigorous formulations of May^{1,2}. Our analysis differs from May's contributions in five important ways. First, we use a measure of interaction strength that is based upon empirical estimates of per capita interaction strength; second, we assume that communities can display nonequilibrium dynamics; third, we construct complexity as simple food webs (after ref. 17) in a manner consistent with patterns found in nature^{14–16}; fourth, we use biomass as the model currency; and fifth, we use consumption rates that become saturated as resource density increases (that is, we use type II functional responses). We describe our model and define terms in Box 1.

It is well known that the model food chain (Fig. 1a) exhibits several behaviours (such as stable equilibria, cycles, chaos and multiple attractors)^{18–20}. In a simplified sense, the food chain is best understood by considering it as two coupled consumer–resource subsystems: a consumer–resource interaction (that is, the interaction between C_1 and R in Equation set (1), box 1) and the top-predator–consumer interaction (that is, the interaction between P and C_1 of Equation set (1)). For example, if the food chain is constructed from two strong consumer–resource interactions (that is, both subsystems produce cyclic behaviour) then the food chain behaviour becomes quite complex and variable^{18–20}. In this case, the food chain can be seen two coupled oscillators whose dynamics depend on whether the frequencies of these oscillators are commensurate (producing cyclic dynamics) or incommensurate (producing quasi-periodic or chaotic dynamics).

Two corollaries follow from these statements: first, stabilizing all the underlying oscillators eliminates the occurrence of cyclic or chaotic dynamics in the full system; and second, reducing the amplitude of the underlying oscillators reduces the amplitude of the dynamics of the full system. Thus we predict that inhibiting strong consumer–resource interactions within a food web promotes persistence in food webs.

We study below how interaction strength influences oscillatory subsystems of more complicated food webs (Fig. 1b–f). We show that weak interactions can act to inhibit potentially oscillatory

subsystems through the following three naturally occurring mechanisms.

In the apparent competition mechanism, apparent competition among resource species occurs when a consumer preys upon multiple resources. In this case, a consumer can inhibit a potentially oscillatory consumer–resource interaction when it trades off preference for one resource (that is, that resource involved in the potentially oscillatory consumer–resource interaction) in order to feed on a second resource. This effectively reduces the efficiency at which the consumer attacks the first resource.

In the exploitative competition mechanism, two consumers compete for the same resource. In this case, the addition of a second competitor reduces the growth rate of the shared resource item (from the perspective of the first consumer) and, therefore, can inhibit a potentially oscillatory consumer–resource interaction involving the first consumer.

In the food-chain-predation mechanism, food-chain predation occurs when a top predator feeds on an intermediate consumer which feeds on a resource. The top predator can inhibit the growth rate of its resource (the intermediate consumer) and, therefore, inhibit the intermediate-consumer–resource trophic interaction. The top predator reduces the intermediate consumer's attack rate on the resource item.

At the heart of these mechanisms is the concept that a stable consumer–resource interaction is required to dampen the dynamic

Box1 The food-web model

We used the following model:

$$\begin{aligned} \frac{dR}{dt} &= R \left(1 - \frac{R}{K} \right) - \frac{\Omega_{C_1 R} x_{C_1} y_{C_1} C_1 R}{\Omega_{C_1 R} R + (1 - \Omega_{C_1 R}) C_2 + R_0} - \frac{\Omega_{C_2 R} x_{C_2} y_{C_2} C_2 R}{R + R_0} \\ \frac{dC_1}{dt} &= -x_{C_1} C_1 \left(1 - \frac{\Omega_{C_1 R} y_{C_1} R + (1 - \Omega_{C_1 R}) y_{C_1} C_2}{\Omega_{C_1 R} R + (1 - \Omega_{C_1 R}) C_2 + R_0} \right) \\ &\quad - \frac{\Omega_{PC_1} x_P y_P C_1 P}{\Omega_{PC_1} C_1 + (1 - \Omega_{PC_1}) C_2 + C_0} \\ \frac{dC_2}{dt} &= -x_{C_2} C_2 \left(1 - \frac{\Omega_{C_2 R} y_{C_2} R}{R + R_0} \right) - \frac{(1 - \Omega_{PC_1}) x_P y_P C_2 P}{\Omega_{PC_1} C_1 + (1 - \Omega_{PC_1}) C_2 + C_0} \\ &\quad - \frac{(1 - \Omega_{C_1 R}) x_{C_1} y_{C_1} C_1 C_2}{\Omega_{C_1 R} R + (1 - \Omega_{C_1 R}) C_2 + R_0} \\ \frac{dP}{dt} &= -x_P P \left(1 - \frac{\Omega_{PC_1} y_P C_1 + (1 - \Omega_{PC_1}) y_P C_2}{\Omega_{PC_1} C_1 + (1 - \Omega_{PC_1}) C_2 + C_0} \right) \end{aligned} \quad (1)$$

where R is resource density, C_1 is the density of the first consumer species, C_2 is density of the second consumer species and P is density of the top predator. The parameters correspond to a bioenergetic interpretation of the Rosenzweig–MacArthur model²⁸, in which K is the resource carrying capacity; R_0 and C_0 are the half-saturation densities of the resource, R , and consumer, C_1 , respectively; x_i is the mass-specific metabolic rate of species i , measured relative to the production-to-biomass ratio of the resource population; y_i is a measure of ingestion rate per unit metabolic rate of species i ; and Ω_{ij} is a fraction indicating the preference of species i for consuming resource species j (ref. 29). Through choice of preference parameters (Ω_{ij}), this model can be made to represent the simple food-chain model (Fig. 1a), exploitative competition (Fig. 1b), apparent competition (Fig. 1c) and intraguild predation (Fig. 1d).

Equation set (1) relies on the assumption of saturating consumption rate, $F_k(i, j)$, such that a consumer, k , captures resources i and j according to the following type II multispecies functional response²⁹.

$$\begin{aligned} F_k(i, j) &= F_k(i) + F_k(j) = \frac{\Omega_{ki} x_k y_k i}{\Omega_{ki} i + (1 - \Omega_{ki}) j + i_0} + \frac{(1 - \Omega_{ki}) x_k y_k j}{\Omega_{ki} i + (1 - \Omega_{ki}) j + i_0} \\ &= \frac{\Omega_{ki} x_k y_k i + (1 - \Omega_{ki}) x_k y_k j}{\Omega_{ki} i + (1 - \Omega_{ki}) j + i_0} \end{aligned} \quad (2)$$

Equation (2) indicates the flow of biomass from resource i and j to consumer k , so it is natural to define the per capita interaction strength, α_{ki} , of resource species i on consumer species k as:

$$\alpha_{ki} = \frac{F_k(i)}{i} = \frac{\Omega_{ki} x_k y_k}{\Omega_{ki} i + (1 - \Omega_{ki}) j + i_0} \quad (3)$$

We define the interaction strength, I_{ki} , as the maximum per capita interaction strength of resource species j on the consumer species i . As α_{ki} has its maximum when resource densities approach zero (that is, when all consumer i 's resources approach 0), then the interaction strength is:

$$I_{ki} = \frac{\Omega_{ki} x_k y_k}{i_0} \quad (4)$$

If we assume a type I functional response ($F(j) = \Omega_{ki} x_k y_k j$) then Equation (3) is the standard form for the per capita interaction strength⁷⁹. Similarly, the interaction strength of consumer species i on resource species j can be defined to be equal in magnitude to Equation (3) and Equation (4) but opposite in sign ($\alpha_{ji} = -\alpha_{ij}$; $I_{ji} = -I_{ij}$).

We now estimate²⁸ the maximum biologically plausible value of interaction strength I_{ij} in system (1). If we know average adult body sizes and the metabolic type (endotherm, vertebrate ectotherm or invertebrate) of species i , we can then estimate biologically plausible values for x_i and for the maximum value of y_i ($y_{\max, i}$; this occurs when ingestion, y_i , is limited by an animal's physiological capacity). Hence, given i 's preference for species j and the half-saturation density, i_0 , we can estimate the interaction strength, which we call the interaction scope as it defines the maximum biologically plausible upper limit to I_{ij} . In nature it is likely that consumers operate at a maximum that is some fraction of the interaction scope (at the realizable interaction scope) because of limitation by ecological factors²⁸. Nevertheless, the estimates for interaction scope allow us to make some predictions concerning food-web complexity and metabolic type.

Assume that any consuming species involved in a single consumer–resource interaction is operating at its realizable interaction scope (that is, at its realizable maximum rate of ingestion). Hence, as species i is operating at its realizable interaction scope, while consuming a single species j (that is, $\Omega_{ij} = 1$), then choosing to consume another resource item requires that species i trade off some of its preference for consuming species j ($\Omega_{ij} < 1$), so the preference for the new resource is $1 - \Omega_{ij}$ (ref. 29). This allows us to compare complexity–stability arguments in a consistent manner.

behaviour of a potentially strong interaction (and, hence, a potentially oscillatory interaction). Consistent with the theory of consumer–resource interactions, reduction in growth rates of the resource and reduction in attack rates by the consumer tend to stabilize a consumer–resource interaction²¹.

We now blend more food-web structure into the simple food-chain model (Fig. 1a), enabling us to study the dynamic implications of naturally occurring food-web structures^{4,14–16}. Looking at exploitative competition²² (Fig. 1b), apparent competition²³ (Fig. 1c), and intraguild predation²⁴ in which consumer 1 feeds on both the basal resource and a second consumer (Fig. 1d), we discuss how the three inhibiting mechanisms naturally arise and work to stabilize food-web dynamics. We then discuss how these mechanisms also arose in previous studies in which it was found that weak amounts of omnivory^{22,25,26} (Fig. 1e) and external (allochthonous) inputs^{4,27} (Fig. 1f) can enhance food-web stability.

To study exploitative competition, we began with parameters for a biologically plausible example of persistent chaotic dynamics for a three-species food chain (Ω_{ij} values = 1; x_{C_1} = 0.40; y_{C_1} = 2.009; x_P = 0.08; y_P = 5.0; C_0 = 0.50; and R_0 = 0.16129; see Box 1 for definitions)¹⁹. Figure 1b shows the addition to this chain of a second intermediate consumer (C_2), which interacts with the basal resource, R , with a strength governed by the parameter Ω_{C_2R} . In this case, Equation set (1) has the following extra parameters: x_{C_2} = 0.20; y_{C_2} = 3.50; R_{02} = 0.90. We chose the new consumer, C_2 , to be competitively inferior to C_1 , so its ability to persist is mediated by the selective predation of the top predator, P , on C_1 . Figure 2a depicts the local minima and maxima attained for the top predator densities, P , in solutions to Equation set (1) across a range of C_2 – R interaction strengths relative to C_1 – R interaction strengths (that is, I_{C_2R}/I_{C_1R}). Figure 2a also shows food-web diagrams that depict the change in food-web structure as the relative interaction strength changes value.

In this scenario we expect the exploitative competition mechanism

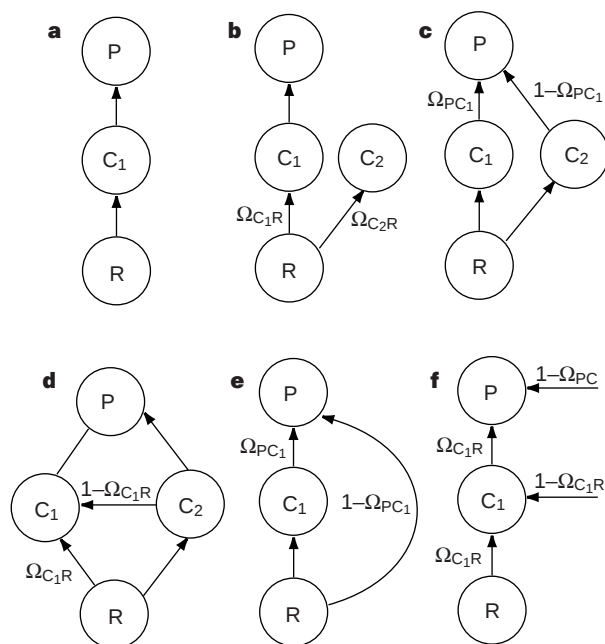


Figure 1 The six food-web configurations studied are: **a**, a simple food chain, **b**, a food web with multiple intermediate consumers (exploitative competition), **c**, a food web with the top predator feeding on two intermediate consumers (apparent competition), **d**, a food web with consumer 1 feeding on the basal resource and on the second consumer (intraguild predation), **e**, a food chain including omnivory, and **f**, a food chain with external inputs. R denotes the basal resource species; C_1 and C_2 denote intermediate consumer species 1 and 2; and P denotes the top predator (Box 1).

ism to inhibit the oscillating C_1 – R subsystem. It does exactly this when C_2 can invade ($I_{C_2R}/I_{C_1R} \approx 0.102$). Below this value the original food chain (P – C_1 – R) remains intact and continues to exhibit chaotic dynamics (shown by the thick set of points from 0.50 to 0.57 in Fig. 2a). However, when C_2 can invade, the dynamics immediately begin to take on a much simpler periodic signal that tends further away from zero (0.102–0.125). Equation set (1) does not reach an equilibrium solution over this range, as the P – C_1 oscillator remains intact—none of the mechanisms is operating to inhibit this oscillatory component.

When the relative interaction strength increases beyond 0.125, the attractor begins to become less bounded again and eventually undergoes a period-doubling cascade to a more complex dynamical regime above $I_{C_2R}/I_{C_1R} \approx 0.15$. The new C_2 – R interaction has become too strong (and therefore oscillatory) and no longer dampens the system; it contributes to the chaos by adding a third oscillating subsystem to the food web.

To study apparent competition, we began with the same parameters as above except that we let $I_{C_2R}/I_{C_1R} \approx 0.154$ (that is, $\Omega_{C_2R} = 0.98$). This returns us to a complex oscillatory food-web dynamic (Fig. 2a); however, in this case, we have three oscillatory subsystems (P – C_1 , C_1 – R , and C_2 – R). We now construct a link between the top predator, P , and the second intermediate consumer, C_2 (Fig. 1c), by allowing Ω_{PC_2} to be < 1 , creating apparent competition between the two intermediate consumers.

Figure 2b shows the local minima and maxima for the top predator, P , in solutions to Equation set (1) across a range of strengths of the P – C_2 interaction relative to strengths of the P – C_1 interaction (I_{PC_2}/I_{PC_1}). We expect the apparent competition mechanism to inhibit the C_2 – R and C_1 – R oscillators. As relative interaction strength increases from zero, it immediately causes a period-doubling reversal, forcing simpler, more bounded, limit cycle dynamics for $I_{PC_2}/I_{PC_1} \approx 0.03$ –0.10 and stable equilibrium dynamics for $I_{PC_2}/I_{PC_1} \approx 0.10$ –0.12. As before, increasing the strength of the interaction further eventually destabilizes the system. Despite the complexity of this system, which includes multiple attractors and numerous bifurcations, qualitatively the result remains: relatively weak links simplify and bound the dynamics of food webs.

We now construct a link such that C_1 can feed on C_2 (Fig. 1d), creating intraguild predation within the food web. The parameters are the same as in the previous case, with $I_{PC_2}/I_{PC_1} \approx 0.01$; (that is, $\Omega_{PC_2} = 0.99$) so that we start off with a complex oscillatory dynamic (we have three oscillators in the food web again). Figure 2c shows the local minima and maxima for the top predator, P , in solutions to Equation set (1) across a range of the relative interaction strength $I_{C_1C_2}/I_{C_1R}$. Here, we expect the apparent competition mechanism to inhibit the C_2 – R subsystem and we expect the food-chain mechanism to inhibit the C_1 – R subsystem. As a result of having two inhibitors and three potential oscillators, the dynamics never reach a locally stable equilibrium; they attain a well-bounded limit cycle solution (driven by the remaining oscillating P – C_1 subsystem) for most $I_{C_1C_2}/I_{C_1R}$ values above 0.05.

In two of the cases above, weak links failed to beget stable local equilibria with all species present—at best we get well-bounded limit cycle solutions. This is because, in these two cases, there is not an inhibiting mechanism for each potential oscillator. This is largely a result of our choice of starting with chaotic dynamics and adding only one additional food-web interaction at a time: our analysis biases our results to have fewer inhibitors than oscillators. Finally we show that adding another, appropriately directed, inhibiting mechanism to such a situation allows for rapid local stabilization to an equilibrium solution. Figure 2d depicts solutions to system 1 when we start off with $I_{C_2R}/I_{C_1R} = 0.11$ (Fig. 2a). Thus, we are starting off with one oscillating subsystem (the P – C_1 subsystem). After adding the apparent competition mechanism, which inhibits the P – C_1 subsystem, we rapidly get a locally stable solution for weak relative interaction strengths (by $I_{PC_2}/I_{PC_1} \approx 0.040$ in Fig. 2d).

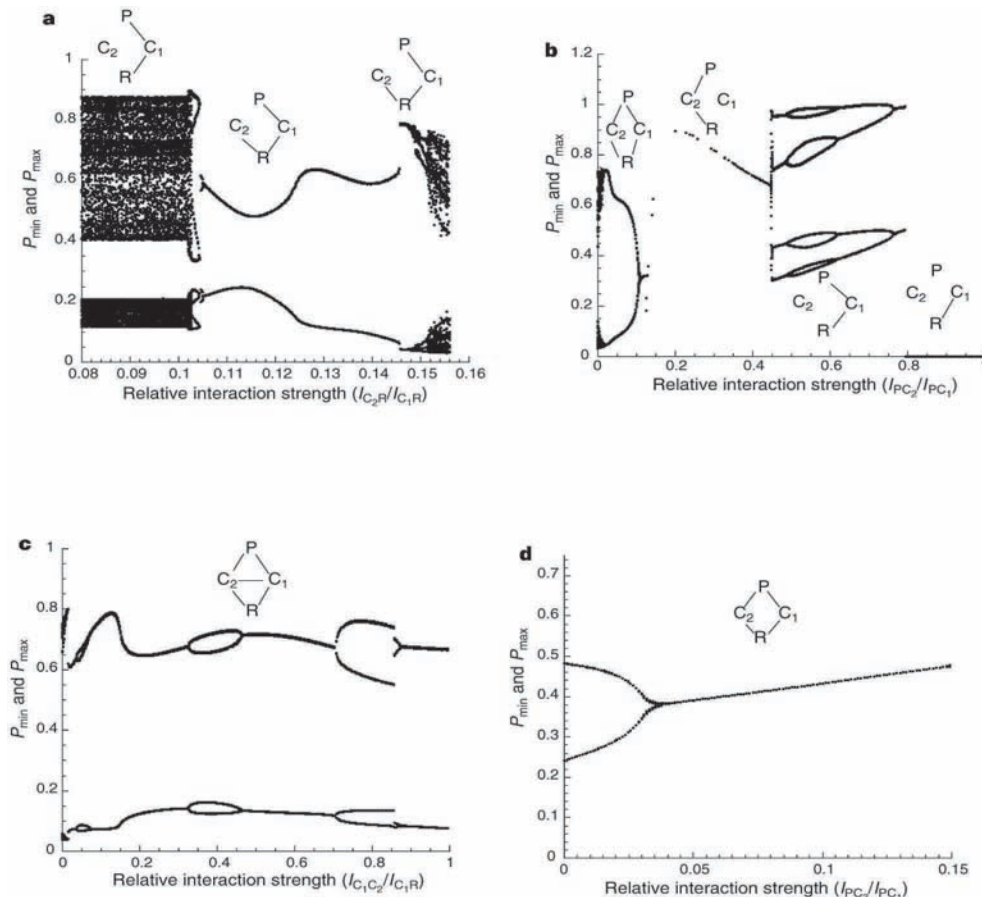


Figure 2 The local minima and maxima for top predator density, P , attained in the attracting solutions for a range of relative interaction strengths. Food-web configurations are given as a function of the relative interaction strengths. Food-web configurations are given as a function of the relative interaction strength.

Using a model formulation similar to system 1, two groups^{26,27} have shown that weak amounts of omnivory (Fig. 1e) and allochthonous inputs (Fig. 1f) bound the magnitude of the oscillations in P – C_1 – R densities further away from zero. In both cases, chaotic dynamics collapsed through period-doubling reversals (that is, the periodicity of solutions reduced from n to $n/2$ to $n/4$ etc.) into well-bounded cyclic or stable dynamics under weak to moderate amounts of omnivory and allochthonous inputs. Both results can be explained with our proposed mechanisms.

Several testable predictions for food webs result from our study. First, if a relatively weak interaction exists for each strong consumer–resource interaction, then the food web should be stabilized relative to the oscillatory subsystems (that is, the food web should be less oscillatory). Second, if food webs have many weak interactions, then, at least from a deterministic viewpoint, chaotic dynamics are unlikely. Third, generalist-dominated food webs should exhibit less variable dynamics than specialist-dominated food webs. Fourth, depauperate food webs should tend to be more oscillatory than reticulate food webs as depauperate food web species tend to have larger average interaction strengths, thus promoting the dominance of a few strong (oscillatory) interactions. Finally, if we assume the realizable interaction scope is proportional to interaction scope, then given all else equal, endotherms ($y_{max,i} \approx 1:60$) and vertebrate ectotherms ($y_{max,i} \approx 3:90$) are more likely to be stabilized by weak food web links than invertebrates as invertebrates have a greater interaction scope (since $y_{max,i} \approx 19:4$) and thus greater potential to maintain a larger number of strong consumer–resource interactions. A larger number of strong

consumer–resource interactions requires a greater number of weak interactions to inhibit oscillatory subsystems. It follows, given topologically identical food webs, that invertebrate-dominated communities are more likely to have the most oscillatory dynamics. Our overall conclusion is that knowledge of interaction strength in study of food webs is vital. Although few quantitative field estimates of interaction strength are available, early data unequivocally indicate that distributions of interaction strength are strongly skewed towards weak interactions^{7–11}. It seems, then, that weak interactions may be the glue that binds natural communities together.

M

Received 17 June; accepted 15 August 1998.

- May, R. M. Stability in multi-species community models. *Math. Biosci.* **12**, 59–79 (1971).
- May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, Princeton, 1973).
- Winemiller, K. O. Spatial and temporal variation in tropical fish trophic networks. *Ecol. Monogr.* **60**, 331–367 (1990).
- Polis, G. A. Complex trophic interactions in deserts: an empirical critique of food web theory. *Am. Nat.* **138**, 123–155 (1991).
- Lavigne, D. M. in *Studies of High Latitude Homeotherms in Cold Ocean Systems* (ed. Montevecchi, W. A.) 95–102 (Canadian Wildlife Service Occasional Paper, Ottawa, Canada, 1995).
- Gardner, M. & Ashby, W. R. Connectance of large dynamical (cybernetic) systems: critical value for stability. *Nature* **228**, 784 (1970).
- Paine, R. T. Food-web analysis through field measurement of per capita interaction strength. *Nature* **355**, 73–75 (1992).
- Fagan, W. F. & Hurd, L. E. Hatch density variation of a generalist arthropod predator: population consequences and community impact. *Ecology* **75**, 2022–2032 (1994).
- Wootton, J. T. Estimates and test of per capita interaction strength: diet, abundance, and impact of intertidally foraging birds. *Ecol. Monogr.* **67**, 45–64 (1997).
- Goldwasser, L. & Roughgarden, J. Construction and analysis of a large Caribbean food web. *Ecology* **74**, 1216–1233 (1993).
- Rafaeli, D. G. & Hall, S. J. in *Food Webs: Integration of Patterns & Dynamics* (eds Polis, G. A. & Winemiller, K. O.) 185–191 (Chapman & Hall, New York, 1996).

12. MacArthur, R. H. Fluctuations of animal populations and a measure of community stability. *Ecology* **36**, 533–536 (1955).
13. Elton, C. S. *Ecology of Invasions by Animals and Plants* (Chapman & Hall, London, 1958).
14. Strong, D. Are trophic cascades all wet? Differentiation and donor-control in speciose ecosystems. *Ecology* **73**, 747–754 (1992).
15. Polis, G. A. Food webs, trophic cascades and community structure. *Aust. J. Ecol.* **19**, 121–136 (1994).
16. Polis, G. A. & Strong, D. Food web complexity and community dynamics. *Am. Nat.* **147**, 813–846 (1996).
17. Holt, R. D. In *Multitrophic Interactions* (eds Begon, M., Gange, A. & Brown, V.) 333–350 (Chapman & Hall, London, 1996).
18. Hastings, A. & Powell, T. Chaos in a three species food chain model. *Ecology* **72**, 896–903 (1991).
19. McCann, K. & Yodzis, P. Biological conditions for chaos in a three-species food chain. *Ecology* **75**, 561–564 (1995).
20. De Feo, O. & Rinaldi, S. Singular homoclinic bifurcations in tritrophic food chains. *Math. Biosci.* **148**, 7–20 (1998).
21. Rosenzweig, M. & MacArthur, R. H. Graphical representation and stability conditions of predator-prey interactions. *Am. Nat.* **107**, 275–294 (1963).
22. Diehl, S. Relative consumer sizes and the strengths of direct and indirect interactions in omnivorous feeding relationships. *Oikos* **68**, 151–157 (1993).
23. Holt, R. D. Predation, apparent competition, and the structure of prey communities. *Theor. Popul. Biol.* **12**, 197–229 (1977).
24. Polis, G. A. & Holt, R. D. Intraguild predation: the dynamics of complex trophic interactions. *Tree* **7**, 151–155 (1992).
25. Fagan, W. F. Omnivory as a stabilizing feature of natural communities. *Am. Nat.* **150**, 554–567 (1997).
26. McCann, K. & Hastings, A. Re-evaluating the omnivory-stability relationship in food webs. *Proc. R. Soc. Lond. B* **264**, 1249–1254 (1997).
27. Huxel, G. & McCann, K. Food web stability: the influence of trophic flows across habitats. *Am. Nat.* **152**, 460–469 (1998).
28. Yodzis, P. & Innes, S. Body-size and consumer-resource dynamics. *Am. Nat.* **139**, 1151–1175 (1992).
29. Chesson, J. The estimation and analysis of preference and its relationship to foraging models. *Ecology* **64**, 1297–1304 (1983).

Acknowledgements. We thank P. DeValpine, M. Hoopes, D. Strong, P. Yodzis and J. Paloheimo for discussions; D. Post, M. E. Connors and D. S. Goldberg for a preprint of a related manuscript; and the US National Science Foundation and the Institute of Theoretical Dynamics for their support.

Correspondence and requests for materials should be addressed to K.M. (e-mail: kevin@six.ucdavis.edu).

A jitter after-effect reveals motion-based stabilization of vision

Ikuya Murakami & Patrick Cavanagh

Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, Massachusetts 02138, USA

A shaky hand holding a video camera invariably turns a treasured moment into an annoying, jittery moment. More recent consumer cameras thoughtfully offer stabilization mechanisms to compensate for our unsteady grip. Our eyes face a similar challenge in that they are constantly making small movements even when we try to maintain a fixed gaze¹. What should be substantial, distracting jitter passes completely unseen. Position changes from large eye movements (saccades) seem to be corrected on the basis of extraretinal signals such as the motor commands sent to the eye muscle^{2–5}, and the resulting motion responses seem to be simply switched off^{6,7}. But this approach is impracticable for incessant, small displacements, and here we describe a novel visual illusion that reveals a compensation mechanism based on visual motion signals. Observers were adapted to a patch of dynamic random noise and then viewed a larger pattern of static random noise. The static noise in the unadapted regions then appeared to ‘jitter’ coherently in random directions. Several observations indicate that this visual jitter directly reflects fixational eye movements. We propose a model that accounts for this illusion as well as the stability of the visual world during small and/or slow eye movements such as fixational drift, smooth pursuit and low-amplitude mechanical vibrations of the eyes.

The experimental setting required for this illusion has three conditions: (1) adaptation to dynamic random noise in a local region (referred to as the adapted area) for at least several seconds; (2) a successive test with static random noise in the adapted area

plus static noise in a region somewhere near the adapted area (referred to as the unadapted area); and (3) maintained fixation throughout these two periods. During the adaptation period, static noise is typically presented in the unadapted region (Fig. 1a), although leaving it blank does not change the outcome (Fig. 1b). After adaptation, static noise presented in the unadapted region seems to jitter rigidly (all dots moving together) in random directions for several seconds. In contrast, the static noise in the

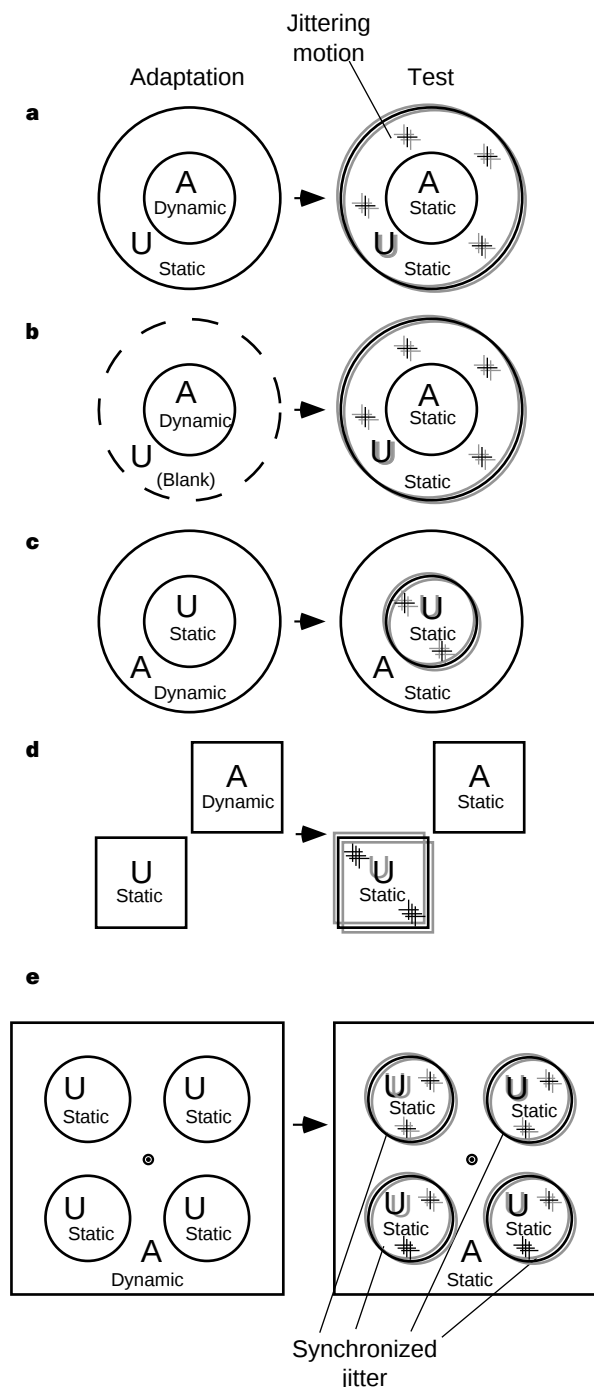


Figure 1 A schematic view of the stimulus configuration and perception. **a–e**, The various configurations used in these experiments and their outcomes. A fixation point is typically provided at the centre of the stimulus, but the illusion occurs for peripheral viewing as well. A and U stand for adapted and unadapted (static or blank) regions, respectively. The blur of circles and crosses in the test stimuli depicts the visual jitter schematically. See <http://visionlab.harvard.edu/> for demonstrations.

adapted region seems stationary (as it actually is). Also, the unadapted region seems to jitter whether the adapted area surrounds the unadapted area or vice versa (Fig. 1c) or whether they are somewhat separated (Fig. 1d). These properties of the jitter were confirmed with tens of naive observers.

A similar adaptation procedure has been reported to produce a distinctly different aftereffect when a blank test field is used. In these experiments, a small uniform region was surrounded by dynamic random noise during adaptation, after which the whole display was turned to a uniform field; the small, unadapted region then seemed to be filled in with dynamic 'twinkle'⁸⁻¹¹. The twinkle was an asynchronous, fine-grained scintillation with no sense of the common motion (jitter) seen on our static patterned test. Moreover, no 'twinkle' aftereffects could be observed in our displays when a uniform test field was used, probably because the unadapted regions in our test were slightly larger, and the adapted regions were much smaller, than those described in the 'twinkle' experiments.

We now describe several characteristics of the visual jitter effect. The duration of visual jitter was measured for the configurations shown in Fig. 1a, b for three conditions. The subjects observed the adaptation and test stimuli monocularly with either the same eye in both periods or different eyes, or binocularly in both periods. When the adaptation and test were in the same eye ('intraocular' and 'binocular' conditions), visual jitter lasted for approximately 5–12 s (Fig. 2). But when the adaptation and test were in different eyes ('interocular' condition), there was no transfer of the effect, unlike the substantial, although incomplete, transfer found for conventional motion aftereffects¹². This result indicates that the adaptation is purely monocular and probably occurs at a very early level.

As motion adaptation in one region has been reported to induce a motion aftereffect in adjacent regions as well¹³⁻¹⁵, visual jitter might be a type of induced jitter. However, one result in particular clearly differentiates visual jitter from other types of motion aftereffect: when unadapted areas were placed at four separate locations in the

adapting dynamic noise (Fig. 1e), the moment-to-moment directions of the jitter aftereffect seen in each were synchronized across all four regions. All regions moved together. No current model of the motion aftereffect could account for this synchronization. Rather, a single source of the illusory motion seems more likely, and fixational eye movements during the test period are the most obvious candidate.

If so, there should not be any visual jitter when the test stimulus is stabilized on the retina. We found this to be true. We analysed the subjective rating of jitter as a function of time after adaptation to (non-stabilized) dynamic noise. The magnitude of the jitter decreased gradually when non-stabilized noise was used throughout the test duration (Fig. 3, conditions N–N). However, when tested with retinally stabilized 'static' noise (the afterimage of a static noise field induced by a stroboscopic flash), no jitter was perceived (condition S–S). When the stabilized test was followed 3 s later by non-stabilized static noise, weaker but clear jitter re-emerged (condition S–N). In contrast, the jitter perceived in a non-stabilized image was extinguished when the stimulus was subsequently replaced by a stabilized image (condition N–S).

These characteristics suggest that the retinal slip caused by small eye movements during fixation is somehow unveiled in the unadapted area. However, the standard models of compensation for eye movements all predict that the jitter should be seen, if anywhere, in the adapted area. Efference copy of motor commands³⁻⁵ or proprioceptive signals from extraocular muscles¹⁶⁻¹⁸ are claimed to correct position changes caused by eye movements¹⁹⁻²¹. This would act to cancel the retinal slip in both adapted and unadapted areas. If we assume that appropriate position correction also suppresses motion responses, say by analogously subtracting the expected motion response, the correction should make the unadapted area appear motionless. But it would overcompensate in the adapted areas where the motion responses to the retinal slip have been attenuated (we verified that motion thresholds in, for example,

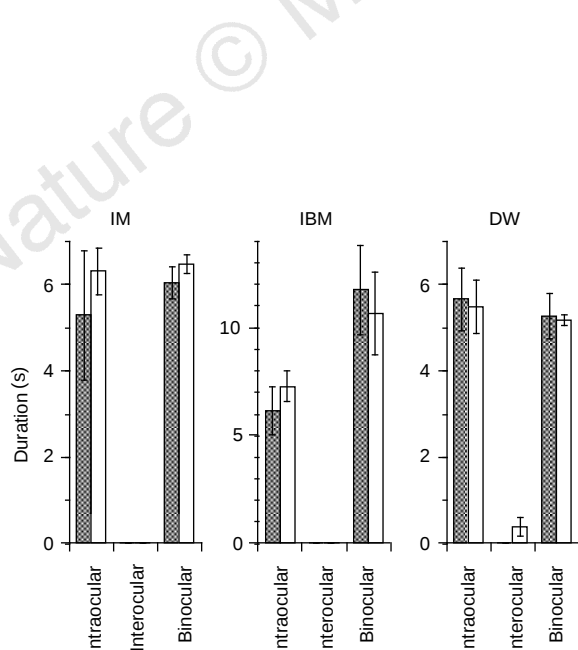


Figure 2 Duration of visual jitter for three subjects, I.M., I.B.M., and D.W. Filled and open bars indicate the results for the stimulus configuration shown in Fig. 1a (static unadapted) and Fig. 1b (blank unadapted), respectively. In the 'intraocular' condition, the same eye (either left or right) was used in both adaptation and test periods. In the 'interocular' condition, different eyes were used for adaptation and test. Both eyes were used in the 'binocular' condition. Each value is the mean of four repeated measurements; the error bars indicate ± 1 s.e.m.

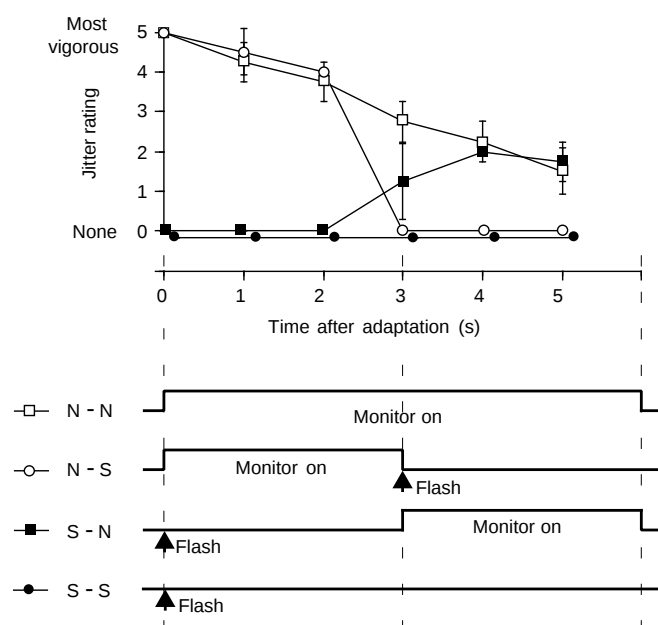


Figure 3 Typical rating data of a naive subject that show no visual jitter in the stabilized retinal image. Each data point is the mean of four repeats; error bars indicate ± 1 s.d. The judged magnitude of jitter is plotted against the time at which the judgement was made, with the beginning of the test period set to time 0. Different symbols indicate different test sequences (N, non-stabilized; S, stabilized). Below the graph the physical appearance and disappearance of the noise in the monitor is shown schematically alongside each symbol. The arrowheads indicate the time at which the flash was presented.

Gabor tests were elevated by adaptation to dynamic noise). Therefore, if any area should jitter, it should be the adapted areas, but the opposite is reported.

Indeed, small eye movements have often been considered to be too small to influence visual functions^{22,23} and, unlike saccades, they are perhaps even too small to be accurately corrected by extraretinal signals. Nevertheless, they are quite capable of driving robust motion responses because motion mechanisms signal displacements at much smaller amplitudes than position mechanisms^{24,25}. Motion responses triggered by saccades can be suppressed over the whole visual field^{6,7}, but this global suppression is not apparent during small eye movements as object motion is easily seen during the incessant small eye movements of normal fixation.

We propose that the motion signals triggered by small, low-velocity eye movements are compensated for solely on the basis of visual motion signals themselves. When one fixates a scene that has both moving and stationary components, the eye-movement velocity is added to the velocities of all image points on the retina. We suggest that a baseline value is recovered from the region of the retina having the lowest instantaneous velocity. The assumption is that this region corresponds to the stationary features of the scene and that motion in this region is due solely to eye movements. Subtract this baseline value from the velocities of all points on the retina and, if the assumption is appropriate, the result will be the desired zero velocity for the stationary regions and the correct velocity for the moving objects. This scheme convincingly explains two questions about visual jitter: (1) why jitter is perceived in the unadapted region and (2) why the adapted region seems stationary. Adaptation to dynamic random noise transiently attenuates responses to motion within the adapted region, creating a new baseline minimum velocity there. Other retinal velocities in the neighbourhood are computed as deviations from this new baseline. In the unadapted region, the unattenuated motion response to eye movements is now above the new, lower baseline. Therefore, the residual, after subtraction, is interpreted as a jittering motion in the unadapted region. In addition, because the adapted region is the source of the new baseline, it appropriately cancels its own motion to result in a zero vector, thereby seeming stationary.

Our proposal requires that the image points that have the current

minimum velocity on the retina can be identified and that their velocity acts as the baseline representing the eye movements. We consider this rule quite plausible. It works well for generic situations and fails only if an object's motion accidentally correlates with eye movements over some period. Another point is that the minimum of retinal velocities is a simple statistical variable that early visual areas might find feasible to extract. Moreover, this approach is superior to simpler statistics such as a global motion average (as effectively computed by large surrounds of opponent-motion cells^{26,27}). Simple averaging cannot differentiate between scene elements that are stationary and those that are moving. This inappropriate correction for motion responses to small eye movements would invariably make stationary parts of the field seem to drift. For example, the interior of a car would seem to rise continuously as we drive forward (the average motion visible through the car windows is downwards for a textured ground and featureless sky).

Is our scheme applied to other kinds of eye movement? Our observations have shown that several types of eye movement during the test period alter the pattern of jitter seen in the manner expected for our model (Fig. 4). For example, we can induce back-and-forth oscillations of gaze by rotating the observer rapidly around the vertical axis (post-rotatory nystagmus). In the interval between a standard adaptation and test, the subject was rotated for several seconds with eyes closed and suddenly stopped. The direction of the visual jitter seen in the unadapted area during the test was almost always horizontal with quick and slow phases, in accordance with the nature of the post-rotatory nystagmus. A similar result held for voluntary smooth pursuits. When, after standard adaptation, a pursuit is made to track a moving spot during test, the unadapted background slides rapidly in the opposite direction. This is the direction of the retinal slip, which is now visible. Furthermore, visual jitter was found to mimic even retinal slip in the eye moved by an external force. After adaptation, up-and-down eye displacements were induced externally by a rapid alternation of extension/relaxation in the nearby skin; at amplitudes of vibration that did not produce noticeable motions of the world, vertical jitter nevertheless occurred in the unadapted area of the static test with a temporal profile synchronized with this vibration. This result suggests that the motion-based correction can compensate for small-amplitude displacements whether driven by the eye muscles or mechanically (such as the vibrations of talking or eating which reach the eyes through bone conduction). Overall, the jitter aftereffect is only one of a wide pattern of motions that can be seen in the unadapted area, each mirroring the eye movements occurring during the test.

The nature of these aftereffects supports the claim that visual motion information is used to correct for the spurious motion signals generated by various eye movements including nystagmus, smooth pursuit and fixational eye movements. Our model complements the extraretinally based processes^{2-5,16-18}, which correct for large position shifts due to saccades but do not affect the motion signals elicited by small eye movements. Without the motion-based process that we propose, one would suffer from an incessant jittering of the visual world owing to the motion responses to small displacements of the eyes.

Note added in proof: We are happy to acknowledge that this jitter effect has been previously observed by Stuart M. Anstis and Richard Gregory (unpublished manuscript). □

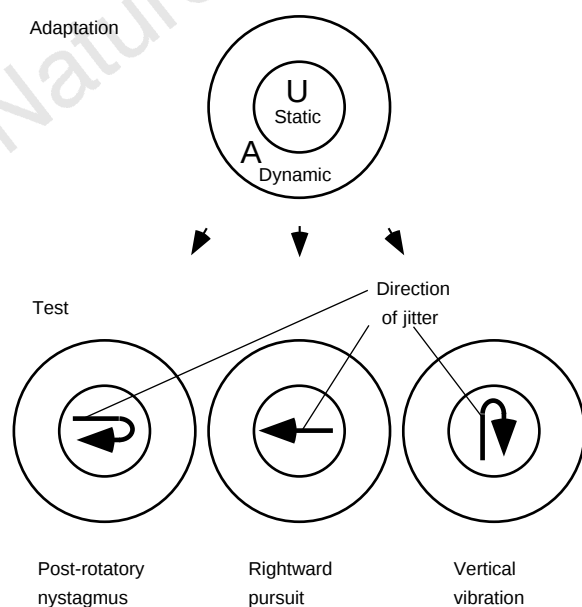


Figure 4 Similar phenomena contingent on various kinds of eye movement during the test period. The black arrows in the central unadapted regions illustrate schematically the direction of illusory motion, which is consistent with the direction of retinal slip due to eye movements indicated below each figure.

Methods

I.M. and two naive subjects participated in formal data acquisition. All had corrected-to-normal visual acuity. In a darkened booth, the stimulus was presented on the screen of a 20-inch colour CRT monitor (Sony GDM2000TC, 832 pixels × 624 pixels) controlled by a computer (Apple Power Macintosh). The subject's head was immobilized with a chinrest; the viewing distance was 77.3 cm.

The adapting stimulus consisted of two concentric regions (see Fig. 1a). A

fixation point was provided in the centre throughout the trial. In the inner circle subtending 6.67° in diameter, black and white dynamic random noise was presented; each dot ($8 \text{ arcmin} \times 8 \text{ arcmin}$) was randomly assigned black (0.18 cd/m^2) or white (52.4 cd/m^2) every frame (75 Hz). In the outer annulus with an outer diameter of 13.33° , static random noise was presented (or not, as in Fig. 1b). There was a uniform grey surround (23.5 cd/m^2) outside the stimuli. After 30 s of adaptation, the two regions were changed to a new pattern of static random noise. To quantify the magnitude of the effect, the total duration of the visual jitter in a 20-s test period was measured; the subject was asked to hold a computer button whenever jitter was apparent and to release the button whenever no jitter was perceived. There was at least 10 s of a blank display between trials.

A stroboscopic flash (Speedotron 4803CX) was used to generate an afterimage under the control of the same computer. It briefly illuminated a printout of the random noise, matched to that on the CRT display, that was otherwise unilluminated and invisible. It was optically superimposed onto the stimulus in the monitor by using a half-silvered mirror. The observation was made binocularly. To make the most conspicuous afterimage, the subject was dark-adapted for at least 10 min before experiments, and the luminance of stimuli in the monitor was reduced (3.1 cd/m^2 for white, 0.01 cd/m^2 otherwise). The stimulus configuration shown in Fig. 1a was used. The monitor was kept totally dark during the period when the subject was to observe the afterimage. Likewise, the perception of afterimage was effectively suppressed when the monitor was turned on again. In this experiment, the method of subjective rating (choosing an integer from 0 to 5, with 0 as no jitter and 5 as the maximum) was used to quantify the magnitude of the illusion because the duration measurement was not appropriate for examining the time course of the effect. In a preliminary experiment, the duration of jitter in the stabilized test (the afterimage) was measured and gave a result of 0 s.

Received 20 April; accepted 19 August 1998.

1. Yarbus, A. L. *Eye Movements and Vision* (Plenum, New York, 1967).
2. Helmholtz, H. von *Handbuch der physiologischen Optik* (Voss, Leipzig, 1866).
3. Skavenski, A. A., Haddad, G. & Steinman, R. M. The extraretinal signal for the visual perception of direction. *Percept. Psychophys.* **11**, 287–290 (1972).
4. Hansen, R. H. & Skavenski, A. A. Accuracy of eye position information for motor control. *Vision Res.* **17**, 919–926 (1977).
5. Matin, L. et al. Oculoparalytic illusion: visual-field dependent spatial mislocalizations by humans partially paralyzed with curare. *Science* **216**, 198–201 (1982).
6. Burr, D., Holt, J., Johnstone, J. R. & Ross, J. Selective depression of motion sensitivity during saccades. *J. Physiol. (Lond.)* **333**, 1–15 (1982).
7. Shioiri, S. & Cavanagh, P. Saccadic suppression of low-level motion. *Vision Res.* **29**, 915–928 (1989).
8. Ramachandran, V. S. & Gregory, R. L. Perceptual filling in of artificially induced scotomas in human vision. *Nature* **350**, 699–702 (1991).
9. Ramachandran, V. S., Gregory, R. L. & Aiken, W. Perceptual fading of visual texture borders. *Vision Res.* **33**, 717–721 (1993).
10. Hardage, L. & Tyler, C. W. Induced twinkle aftereffect as a probe of dynamic visual processing mechanisms. *Vision Res.* **35**, 757–766 (1995).
11. Tyler, C. W. & Hardage, L. Long-range twinkle induction: an achromatic rebound effect in the magnocellular processing system? *Perception* **27**, 203–214 (1998).
12. Barlow, H. B. & Brindley, G. S. Inter-ocular transfer of movement aftereffects during pressure blinding of the stimulated eye. *Nature* **200**, 1347–1347 (1963).
13. Anstis, S. M. & Reinhardt-Rutland, A. H. Interactions between motion aftereffects and induced movement. *Vision Res.* **16**, 1391–1394 (1976).
14. Zaidi, Q. & Sachtler, W. L. Motion adaptation from surrounding stimuli. *Perception* **20**, 703–714 (1991).
15. von Grünau, M. & Dubé, S. Comparing local and remote motion aftereffects. *Spat. Vis.* **6**, 303–314 (1992).
16. Abrahams, V. C. & Anstee, G. Unit activity in the superior colliculus of the cat following passive eye movements. *Can. J. Physiol. Pharmacol.* **57**, 359–368 (1979).
17. Fiorentini, A., Berardi, N. & Maffei, L. Role of extraocular proprioception in the orienting behaviour of cats. *Exp. Brain Res.* **48**, 113–120 (1982).
18. Steinbach, M. Proprioceptive knowledge of eye position. *Vision Res.* **27**, 1737–1744 (1987).
19. Post, R. B. & Leibowitz, H. W. A revised analysis of the role of efference in motion perception. *Exp. Brain Res.* **63**, 631–643 (1985).
20. Duhamel, J. R., Colby, C. L. & Goldberg, M. E. The updating of the representation of visual space in parietal cortex by intended eye-movements. *Science* **255**, 90–92 (1992).
21. Haarmeier, T., Thier, P., Repnow, M. & Petersen, D. False perception of motion in a patient who cannot compensate for eye movements. *Nature* **389**, 849–852 (1997).
22. Sheedy, J. E. The perceived stability of fixation. *Am. J. Optom. Physiol. Opt.* **58**, 149–154 (1981).
23. Badcock, D. R. & Wong, T. L. Resistance to positional noise in human vision. *Nature* **343**, 554–555 (1990).
24. Nakayama, K. & Tyler, C. W. Psychophysical isolation of movement sensitivity by removal of familiar position cues. *Vision Res.* **21**, 427–433 (1981).
25. Seiffert, A. E. & Cavanagh, P. Position displacement, not velocity, is the cue to motion detection of second-order patterns. *Vision Res.* **38**, 3569–3582 (1998).
26. Allman, J., Miezin, F. & McGuinness, E. Direction- and velocity-specific responses from beyond the classical receptive field in the middle temporal visual area (MT). *Perception* **14**, 105–126 (1985).
27. Tanaka, K. et al. Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey. *J. Neurosci.* **6**, 134–144 (1986).

Correspondence and requests for materials should be addressed to I.M. (e-mail: ikuya@wjh.harvard.edu).

The DCC gene product induces apoptosis by a mechanism requiring receptor proteolysis

Patrick Mehlen, Shahrooz Rabizadeh, Scott J. Snipas, Nuria Assa-Munt, Guy S. Salvesen & Dale E. Bredesen

Program on Aging, The Burnham Institute, La Jolla, California 92037, USA
Neuroscience Department, University of California, San Diego, California 92093, USA

The development of colonic carcinoma is associated with the mutation of a specific set of genes¹. One of these, *DCC* (*deleted in colorectal cancer*)^{2–5}, is a candidate tumour-suppressor gene, and encodes a receptor for netrin-1, a molecule involved in axon guidance^{6–8}. Loss of *DCC* expression in tumours is not restricted to colon carcinoma², and, although there is no increase in the frequency of tumour formation in *DCC* hemizygous mice⁵, re-establishment of *DCC* expression suppresses tumorigenicity^{3,4}. However, the mechanism of action of *DCC* is unknown. Here we show that *DCC* induces apoptosis in the absence of ligand binding, but blocks apoptosis when engaged by netrin-1. Furthermore, *DCC* is a caspase substrate, and mutation of the site at which caspase-3 cleaves *DCC* suppresses the pro-apoptotic effect of *DCC* completely. These results indicate that *DCC* may function as a tumour-suppressor protein by inducing apoptosis in settings

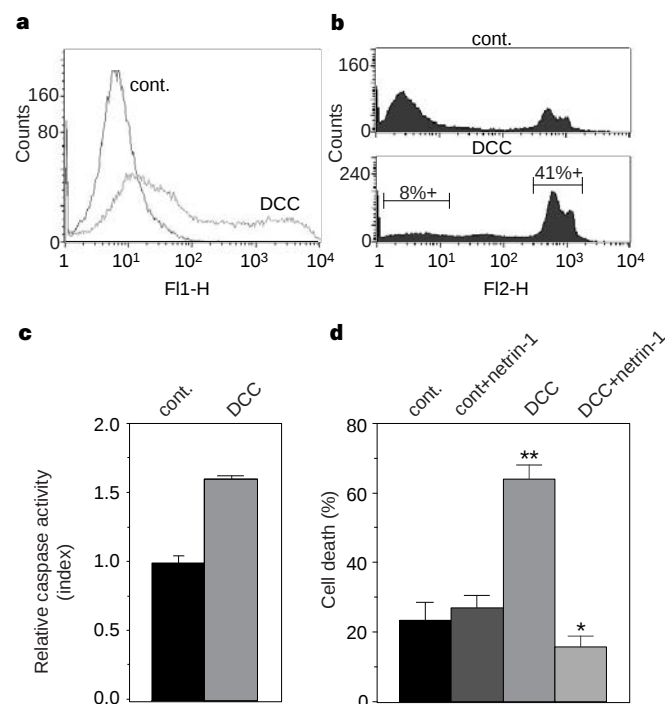


Figure 1 DCC expression induces cell death, which is blocked by netrin-1. 293T cells were transiently transfected with either the DCC expression plasmid pDCC-CMV-S (DCC) or the pCMV control plasmid (cont.). **a**, Expression of DCC 48 h after transfection. **b**, Cell death induced by DCC expression, monitored by flow cytometry using propidium iodide incorporation. Note the increase in dead cells (bottom right peak) after DCC expression. Only 8% of the live cells expressed DCC, in comparison to 41% of the dead cells. **c**, Increase in caspase activity after DCC expression. **d**, Netrin-1 expression blocks DCC-induced cell death. Standard deviations are indicated ($n = 3$). Double asterisk indicates $P < 0.005$ (unpaired t -test); single asterisk indicates $P < 0.05$ (paired t -test).

in which ligand is unavailable (for example, during metastasis or tumour growth beyond local blood supply) through functional caspase cascades by a mechanism that requires cleavage of DCC at Asp 1,290.

Here we show that DCC is a dependence receptor, and therefore functionally related to other dependence receptors such as p75^{NTR} (the common neurotrophin receptor) and the androgen receptor^{9–13}. Such receptors create cellular states of dependence on their respective ligands by inducing apoptosis when unoccupied by ligand, but inhibiting apoptosis in the presence of ligand^{9–12}. We expressed full-length human DCC transiently in 293T human embryonic kidney cells and in CaCO2 colonic carcinoma cells. Figure 1a shows flow cytometric analysis of DCC expression.

To study the involvement of DCC in cell death, we analysed populations transfected with either a mock or a DCC-expressing vector for propidium iodide staining, as an indication of cell death. We detected a marked increase of propidium iodide positive cells in the DCC-transfected population (Fig. 1b). Furthermore, a comparison of live and dead cells in the DCC-transfected population showed an over-representation of DCC expression in the dead cells (Fig. 1b). Hence, expression of DCC leads to cell death in 293T cells. To determine levels of apoptosis, we measured cleavage of Ac-DEVD-AFC as an indication of caspase activation. DCC expression leads to a marked increase in apoptosis (Fig. 1c). Nearly identical results were obtained when DCC was expressed in the CaCO2 cells (results not shown). Taken together, these results indicate that DCC has a pro-apoptotic effect on 293T and CaCO2 cells.

Whereas DCC expression led to apoptosis, co-transfection of netrin-1 with the DCC-expression construct completely suppressed the pro-apoptotic effect of DCC in 293T cells (Fig. 1d). Similar results were obtained using CaCO2 cells (results not shown). The results were the same whether netrin-1 was co-expressed with DCC or simply added to the cell medium in conditioned medium from netrin-1-expressing cells. Netrin-1 expression alone had no effect on cells transfected with a plasmid control, but the expression of both DCC and netrin-1 led to a slight but significant decrease in cell death compared with mock-transfected 293T cells. Hence, expression of

DCC alone induces apoptosis, whereas expression of DCC in the presence of a ligand, netrin-1, has an anti-apoptotic effect.

Because the pro-apoptotic effect of DCC was reversed by netrin-1, we studied the expression of DCC in the presence and absence of netrin-1. Whereas, in the absence of netrin-1, DCC was detected at a low level at 48 h after transfection, the presence of netrin-1 increased the net DCC expression markedly (Fig. 2a). Moreover, a 20-min tamoxifen treatment, which enhances apoptosis induction¹⁴, led to the complete disappearance of DCC on immunoblot (Fig. 2a), but this effect was blocked by netrin-1. Furthermore, the addition of 20 μ M of the cell-penetrant caspase inhibitor zVAD-fmk just before the tamoxifen treatment partially inhibited the disappearance of DCC (Fig. 2a). Similar results were obtained when we expressed the intracellular domain of DCC (DCC-IC): Fig. 2b shows that treatment of transfected cells with 20 μ M zVAD-fmk increased levels of detectable DCC-IC. When the immunoblot was done at an earlier time following transfection (24 h after transfection), we detected specific bands, migrating at relative molecular masses (M_r values) of about 20K and 40K less than the M_r of full-length DCC-IC (Fig. 2c), indicating cleavage approximately 20K and 40K from the carboxy terminus of DCC. The intensity of these two bands was decreased by 4-fold and 19-fold, respectively, in the presence of zVAD-fmk (Fig. 2c). These results indicate that DCC may be proteolytically cleaved when expressed in 293T cells, that this process is blocked by netrin-1, and that the cleavage depends at least partly on caspase activation.

To determine whether DCC is a caspase substrate, we translated DCC-IC using an *in vitro* translation system and ³⁵S-methionine labelling. Translated DCC-IC was incubated with purified active-site-titrated caspases 3, 6, 7 and 8 individually. Incubation with caspase-3 led to the appearance of cleavage products migrating at apparent relative molecular masses of 22K and 20K (Fig. 3a). Incubation with caspases 6 and 8 also led, to a much lesser extent, to a similar cleavage pattern. A minor cleavage product of ~36K was detected when caspase-7 was incubated with DCC-IC, indicating that caspase-7 may cleave DCC-IC at a different site to caspase-3 (Fig. 3a). We mapped the sites of cleavage by caspases by constructing mutants based on preferred P4 (ref. 15) and P1' positions and by

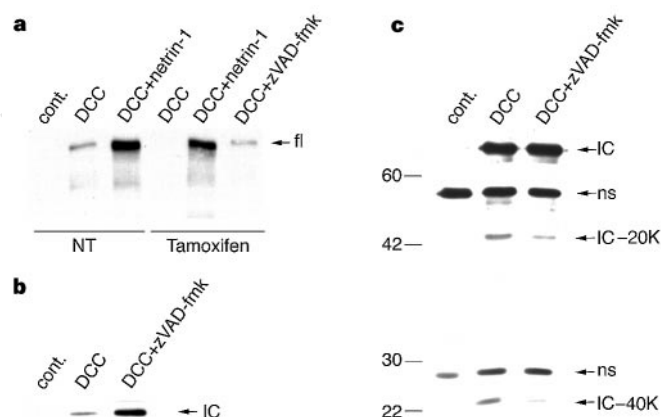


Figure 2 Intracellular processing of DCC and DCC-IC are inhibited by netrin-1 and the caspase inhibitor zVAD-fmk. **a**, Disappearance of DCC expressed in 293T cells. Cells were transfected as in Fig. 1d, and were then left untreated (NT) or treated for 20 min with 17 μ M tamoxifen in the presence or absence of 20 μ M zVAD-fmk. **b**, Increased stability of DCC-IC in the presence of zVAD-fmk (48 h). **c**, DCC-IC is cleaved when expressed in 293T cells (24 h after transfection with DCC-IC). Densitometric quantification showed a 4-fold decrease in the IC-20K product, and 19-fold decrease in the IC-40K product, in the presence of zVAD-fmk. DCC full-length; IC, DCC intracellular domain; ns, nonspecific bands; IC-20K and IC-40K, fragments migrating at a relative molecular mass that is 20K or 40K less than that of DCC-IC.

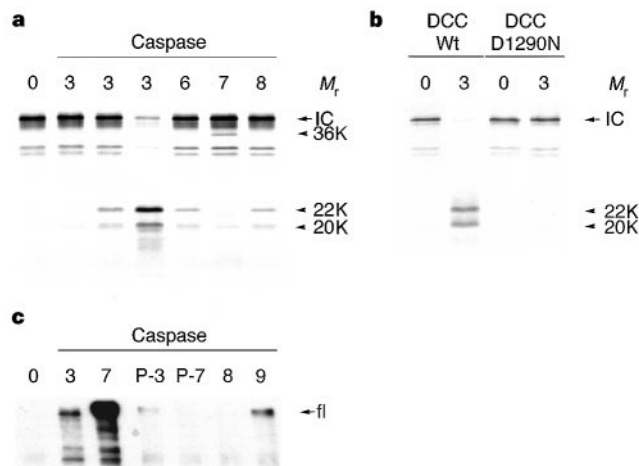


Figure 3 DCC is a caspase substrate *in vitro*. **a**, *In vitro*-translated DCC-IC was incubated in the absence of caspase (first lane) or with purified caspase-3 (0.003 μ M, 0.03 μ M or 0.3 μ M, from left to right), caspase-6 (0.4 μ M), caspase-7 (1.6 μ M), or caspase-8 (0.3 μ M), for 2 h. An autoradiograph is shown. **b**, Asp 1,290 is the major caspase site. The D1290N mutant is not cleaved by caspase-3. **c**, DCC interacts with caspases 3, 7 and 9, but not with caspase-8. Lysates of DCC-transfected 293T cells were incubated in the presence of 0.5 μ M caspase 3, 7, 8 or 9, or with pro-caspase 3 or 7. Immunoprecipitation was then performed as described (see Methods). fl, full-length DCC; wt, wild-type.

using the apparent relative sizes of the *in vivo* cleavage fragments (Fig. 2c). Mutation of Asp 1,290 to Asn in DCC-IC suppressed cleavage by caspase-3 completely (Fig. 3b). We mapped the minor cleavage by caspase-7 to Asp 1,185, using a similar approach (results not shown). Furthermore, immunoprecipitation of caspases 3, 7, 8 and 9, showed that DCC co-immunoprecipitated with caspases 3, 7 and 9, but not with caspase-8 (Fig. 3c).

To determine the function of caspase-mediated cleavage of DCC at the major site, Asp 1,290, we expressed the D1290N mutant of full-length DCC in 293T cells (Fig. 4a), and assessed the level of apoptosis by following caspase activation or trypan blue exclusion. The D1290N mutation completely suppressed the pro-apoptotic effect of DCC, and in fact resulted in a modest but significant anti-apoptotic effect (Fig. 4b, c). In contrast, a control mutant, R1291A, that should not block cleavage by caspases but should affect cleavage by any protease requiring the basic Arg 1,291, had no effect on apoptosis induction by DCC (Fig. 4a–c). Similar results were obtained with stable transfections of 293 cells, using a tetracycline-inducible promoter, and with stable transfection of REGb colonic carcinoma cells and 293T cells, using a constitutive promoter (the cytomegalovirus immediate-early promoter) (Fig. 4d, e). Thus the D1290N mutation converted DCC from a pro-apoptotic receptor (when unoccupied) to an anti-apoptotic (unoccupied) receptor, indicating that caspase-mediated cleavage of DCC is a crucial event in the induction of apoptosis.

We then tested the effects of expressing the caspase cleavage fragments of the DCC-IC (Fig. 5). Expression of the amino-terminal (residues 1,121–1,290), but not C-terminal (residues 1,291–1,447), caspase cleavage fragment of DCC led to apoptosis

induction (Fig. 5). These results, together with the results obtained with the D1290N mutation and from co-expression of netrin-1, indicate that the induction of apoptosis by DCC may require proteolytic cleavage, which occurs primarily for the unoccupied receptor, creating a pro-apoptotic fragment (truncated receptor) that includes the region stretching from residue 1,121 to residue 1,290. Further mutational studies to define the dependence domain (that is, the domain required for apoptosis induction in the absence of ligand) showed that residues 1,243–1,264 are required for apoptosis induction (Fig. 5), whereas residues 1,124–1,209 and 1,210–1,228, two regions of predicted helix formation, are not required for apoptosis induction (results not shown).

These results also indicate that DCC may function as a tumour-suppressor gene (or metastasis suppressor) by inducing apoptosis in cells that are not exposed to ligand, for example, metastatic cells or locally invasive cells. It is also possible that the netrin-1–DCC interaction may block apoptosis during neural development, and it is interesting that netrin-1-null mice lack pontine neuronal nuclei (this may, however, relate to a migration defect rather than an increase in developmental neuronal cell death)⁶. Given the results showing caspase-mediated actin cleavage in both apoptosis and dystrophic neurites¹⁶, one cannot exclude the possibility that the role of DCC in apoptosis may relate biochemically to its role in axon guidance.

The idea that DCC is a dependence receptor is compatible with the association of colorectal cancer with the loss of function of DCC². Our results support the argument that the loss of function of DCC may enhance survival of cells outside the region of ligand availability. Our results are also compatible with the findings that DCC expression in some tumours is associated with increased cell death⁴, that DCC deletion tends to be a late event in the development of colorectal cancer¹⁷, and that DCC deletion tends to be associated with invasive and metastatic tumours¹⁸. It is not yet certain whether netrin-1, which is widely expressed², or another ligand (such as, heparin¹⁹ or an unidentified ligand) may be important in blocking DCC-induced apoptosis *in vivo*.

The expression of two other pro-apoptotic receptors, p75^{NTR} and

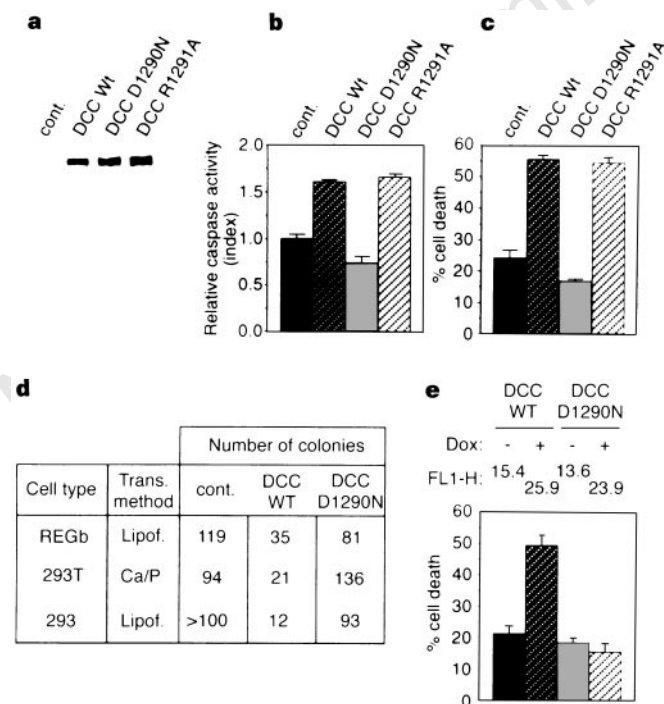


Figure 4 Mutation of the site of cleavage of DCC by caspase-3 suppresses the pro-apoptotic effect of DCC completely. 293T cells were transfected with expression constructs for the indicated mutants of DCC, and were **a**, analysed by immunoblot, or, for determination of cell death, were monitored by **b**, measurement of caspase activation and **c**, **e**, trypan blue exclusion as described in Fig. 1. Standard deviations are indicated ($n = 3$). Each experimental group in **b**, **c**, is significantly different from the control ($P < 0.05$ analysis of variance). **d**, Decrease in colony number in cells stably transfected with pDCC-CMV-S or pTRE-DCC (293). **e**, Cell death induced in 293 cells by doxycycline-inducible expression of DCC (see Methods).

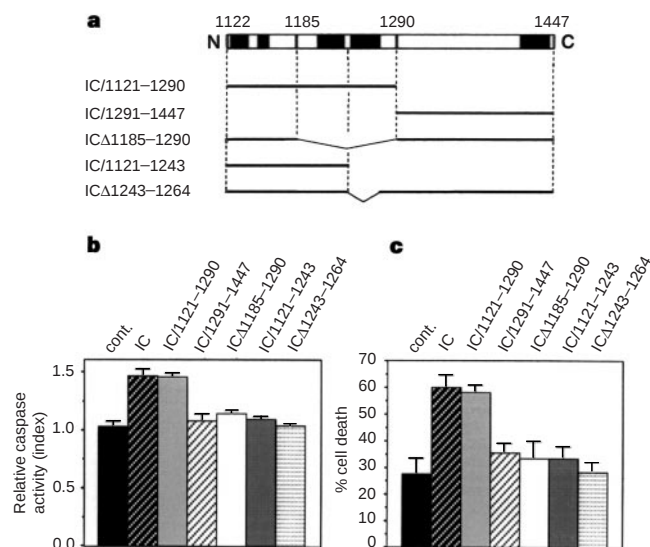


Figure 5 The dependence domain of DCC lies in the N-terminal portion of the intracellular domain. Mutants of the DCC intracellular domain (DCC-IC) were transfected and analysed for apoptosis induction as in Fig. 4. **a**, Map of DCC-IC and the constructs expressed. Black boxes represent the predicted helices. **b**, **c**, Caspase activation (**b**) and cell-death induction (**c**) by DCC-IC mutants. Residues 1,243–1,264 are required for apoptosis induction. Standard deviations are indicated ($n = 3$). IC and IC/1121–1290 are significantly different from control ($P < 0.05$, analysis of variance).

the androgen receptor, tends to be lost late in (prostate) cancer development²⁰. Thus, dependence receptors in general, and DCC in particular, may be viewed as suppressors of metastasis and of locally invasive growth that would exceed ligand availability²¹. □

Methods

Immunoblotting, immunocytochemistry and fluorescence-activated cell sorting (FACS) analysis. One-dimensional immunoblots using antibodies raised against DCC (Oncogene Research Products) or haemagglutinin (HA; Boehringer), were performed as described²². Immunocytochemistry was performed as described²³ and DCC-expressing cells were then monitored using a FAC-sort (Becton Dickinson), with a 488-nm excitation and 530-nm emission filter.

Cell-death analysis. Cell death was analysed using either trypan blue²⁴ or propidium iodide²⁵ staining procedures. Apoptosis was monitored with the ApoAlert caspase assay, which uses the Ac-DEVD-AFC substrate (Clontech). This activity was determined according to the manufacturer's instructions using a Ac-DEVD-CHO-pretreated lysate as a negative control. Caspase activation is shown as the ratio between the caspase activity of the sample and that measured in 293T cells transfected with pCMV; for all samples, the background remaining after inhibition by Ac-DEVD-CHO was subtracted.

Site-directed mutagenesis and plasmid construction. pDCC-CMV-S was obtained from B. Vogelstein, and contains the complementary DNA encoding human DCC under the control of the CMV early promoter²⁶. pGNET1myc was obtained from M. Tessier-Lavigne, and encodes chick netrin-1 (ref. 6).

The plasmid pCR-DCC-IC was constructed by inserting the intracellular domain of DCC into pCR3.1 (Invitrogen) using the polymerase chain reaction (PCR) primers 5'-TGACCATGGGACGAGCTCTTCAGCC-3', and 5'-TGCAAAGTCCCGGAAAATTCACCT-3' with pDCC-CMV-S as template.

pCD-HA-DCC-IC encodes the intracellular domain of DCC linked at its N terminus to a myristoylation signal and an HA epitope. To make this construct, the intracellular domain of DCC was obtained by PCR using pDCC-CMV-S as template and the primers 5'-TGCATATGCGAGCTCTTCAGCCAG-3', and 5'-TGCAAAGTCCCGGAAAATTCACCT-3', and was inserted into pCR3.1. The *EcoRI*–*EcoRI* fragment was then subcloned into pSVK3 (Pharmacia). The *NdeI*–*NdeI* fragment containing the myristoylation signal and an HA epitope²⁷ was then cloned into the previous construct. pCD-HA-DCC-IC was obtained by inserting the *EcoRI*–*EcoRI* fragment of pSVK3-HA-DCC-IC into pCDNA3 (Invitrogen).

The substitutions D → N and R → A at D1290 and R1291, respectively, of DCC were created using the QuickChange site-directed mutagenesis system (Stratagene). pCR-DCC-IC.D1290N, pDCC-CMV.D1290N and, respectively, pCR-DCC-IC.R1291A and pDCC-CMV.R1291A were constructed using pCR-DCC-IC or pDCC-CMV-S as templates and the following primers: 5'-CACTCTCAGTGAACCGAGGTTTC-3' and 5'-GAAACCTCGGTTCACTGAGAGTG-3', or 5'-CACTCTCAGTGGACGCGAGGTTTCGGAGCAG-3', and 5'-CTGCTCCGAAACCTGCGTCCACTGAGAGTG-3'.

pCR-DCC-IC/1121-1290 and pCR-DCC-IC/1290-1447 were obtained as described for pCR-DCC-IC using, respectively, the primers 5'-TGACCATGGGACGAGCTCTTCAGCC-3' and 5'-CTTACACTGAGAGTGTGGGAATGGC-3' or 5'-TGACCATGGGACGAGGTTTCGGAGCA-3' and 5'-TGCAAAGTCCCGGAAAATTCACCT-3'. pCR-DCC-ICΔ1185–1290, pCR-DCC-IC/1121–1243 and pCR-DCC-ICΔ1243–1264 were obtained using the QuickChange site-directed mutagenesis system, pCR-DCC-IC as template and the following primers: 5'-GCAAAGTGGCAAGACCGAGGTTTCGGAGCAG-3' and 5'-CTGCTCCGAAACCTCGGTTCTGGCAACTTTCG-3', 5'-GATGCCCAGTCCAACCTGCTGTCTGAGC-3' and 5'-GCTCAGCAGCAGGTTAGTTGGAGCTGGGCATC-3', or 5'-GATGCCCAGTCCAACCTCCTCCCTGCTCC-3' and 5'-GGGAGACGGGAGGATGTGGACTGGGCATC-3'.

The plasmids pTRE-DCC and pTRE-DCC-D1290N were constructed by inserting cDNA encoding DCC (*XbaI*–*SalI*) or DCC-D1290N into pTRE that was modified by introduction of a *SalI* restriction site into the polylinker.

Cells and transfection procedures. Transient transfections of human embryonic kidney 293T cells were performed as described¹³ according to a modified calcium-phosphate procedure²⁸. Transient transfections of human colon carcinoma CaCO2 cells were performed using lipofectamine and the manufacturer's procedure (Gibco). 293 tet-on cells and the tet-on system were

from Clontech. 293 tet-on cells were co-transfected with pTRE, pTRE-DCC, pTRE-DCC-D1290N and pTK-hyg (ratio 10:1) using lipofectamine, and then selected for hygromycin resistance. DCC expression in 293 tet-on cells was enhanced by addition of 1 μg ml⁻¹ doxycycline for 48 h.

In vitro translation and caspase-mediated cleavage reactions. Plasmids pCR-DCC-IC, and pCR-DCC-IC.D1290N were transcribed using T7 polymerase, then translated using the TNT system (Promega) in the presence of 50 μCi [³⁵S]methionine (Dupont) for 2 h at 30 °C. Translations were incubated for 2 h in 20 mM PIPES, 100 mM NaCl, 1% sucrose, 10 mM dithiothreitol and 0.1 mM EDTA, pH 7.2, at 37 °C in the presence of caspase 3, 6, 7, or 8 as described²⁹.

Caspase-interaction studies. Transfected 293T cells were lysed in 50 mM HEPES 7.6, 250 mM NaCl, 5 mM EDTA and 0.1% NP-40. Cell lysate was then incubated overnight at 4 °C in the presence of 0.5 μM caspase (caspase 3, 7, 8, 9, pro-3 or pro-7). Immunoprecipitation was performed using anti-polyhistidine antibody and protein-A Sepharose (Sigma), and DCC interaction with caspase was monitored by immunoblot using anti-DCC antibody.

Received 10 April; accepted 29 July 1998.

1. Fearon, E. R. *et al.* Identification of a chromosome 18q gene that is altered in colorectal cancers. *Science* **247**, 49–56 (1990).
2. Fearon, E. R. DCC: is there a connection between tumorigenesis and cell guidance molecules? *Biochim. Biophys. Acta* **1288**, M17–M23 (1996).
3. Tanaka, K. *et al.* Suppression of tumorigenicity in human colon carcinoma cells by introduction of normal chromosome 5 or 18. *Nature* **349**, 340–342 (1991).
4. Klingelutz, A. J., Hedrick, L., Cho, K. R. & McDougall, J. K. The DCC gene suppresses the malignant phenotype of transformed human epithelial cells. *Oncogene* **10**, 1581–1586 (1995).
5. Fazeli, A. *et al.* Phenotype of mice lacking functional Deleted in colorectal cancer (*Dcc*) gene. *Nature* **386**, 796–804 (1997).
6. Serafini, T. *et al.* Netrin-1 is required for commissural axon guidance in the developing vertebrate nervous system. *Cell* **87**, 1001–1014 (1996).
7. Kennedy, T. E., Serafini, T., de la Torre, J. R. & Tessier-Lavigne, M. Netrins are diffusible chemotropic factors for commissural axons in the embryonic spinal cord. *Cell* **78**, 425–435 (1994).
8. Keino-Masu, K. *et al.* Deleted in colorectal cancer (*DCC*) encodes a netrin receptor. *Cell* **87**, 175–185 (1996).
9. Rabizadeh, S. *et al.* Induction of apoptosis by the low-affinity NGF receptor. *Science* **261**, 345–348 (1993).
10. Rabizadeh, S. & Bredesen, D. E. Is p75^{NTR} involved in developmental neural cell death? *Dev. Neurosci.* **16**, 207–211 (1994).
11. Bredesen, D. E. & Rabizadeh, S. p75^{NTR} and apoptosis: Trk-dependent and Trk-independent effects. *Trends Neurosci.* **20**, 287–290 (1997).
12. Bredesen, D. E. *et al.* p75^{NTR} and the concept of cellular dependence: seeing how the other half die. *Cell Death Differ.* **5**, 365–371 (1998).
13. Ellerby, L. M. *et al.* Kennedy's disease: caspase cleavage of the androgen receptor is a crucial event in cytotoxicity. *J. Neurochem.* (in the press).
14. Ellerby, H. M. *et al.* Establishment of a cell-free system for neuronal apoptosis: comparison of pre-mitochondrial, mitochondrial, and post-mitochondrial phases. *J. Neurosci.* **17**, 6165–6178 (1997).
15. Thornberry, N. A. *et al.* A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *J. Biol. Chem.* **272**, 17907–17911 (1997).
16. Yang, F. *et al.* Antibody to caspase-cleaved actin detects apoptosis in differentiated neuroblastoma and plaque-associated neurons and microglia in Alzheimer's disease. *Am. J. Pathol.* **152**, 379–389 (1998).
17. Cho, K. R. & Fearon, E. R. DCC: linking tumor suppressor genes and altered cell surface interactions in cancer. *Curr. Opin. Genet. Dev.* **5**, 72–78 (1995).
18. Kern, S. E. *et al.* Allelic loss in colorectal carcinomas. *J. Am. Med. Assoc.* **261**, 3099–3103 (1989).
19. Bennett, K. L. *et al.* Deleted in colorectal carcinoma (*DCC*) binds heparin via its fifth fibronectin type III domain. *J. Biol. Chem.* **272**, 26940–26946 (1997).
20. Pflug, B. R., Onoda, M., Lynch, J. H. & Djakiew, D. Reduced expression of the low affinity nerve growth factor receptor in benign and malignant human prostate tissue and loss of expression in four human metastatic prostate tumor cell lines. *Cancer Res.* **52**, 5403–5406 (1992).
21. Bredesen, D. E. in *Apoptosis II: The Molecular Basis of Apoptosis in Disease* 397–421 (eds Tomei, L. D. & Cope, F. O.) (Cold Spring Harb. Lab. Press, Plainview, New York, 1994).
22. Mehlen, P., Kretz Remy, C., Previle, X. & Arrigo, A. -P. Human hsp27, *Drosophila* hsp27 and human αB-crystallin expression-mediated increase in glutathione is essential for the protective activity of these proteins against TNFα-induced cell death. *EMBO J.* **15**, 2695–2706 (1996).
23. Ruan, Y., Camerini, D., Rabizadeh, S. & Bredesen, D. E. Expression of CD40 induces neural apoptosis. *J. Neurosci. Res.* **50**, 383–390 (1997).
24. Mehlen, P. *et al.* Constitutive expression of either human hsp27, *Drosophila* hsp27 or human αB-crystallin confers resistance to tumor necrosis factor- and oxidative stress-induced cytotoxicity in stably transfected murine L929 fibroblasts. *J. Immunol.* **154**, 363–374 (1995).
25. Mehlen, P., Coronas, V., Ljubic, V., Jourdan, F. & Arrigo, A. -P. Small stress protein Hsp27 accumulation during dopamine-mediated differentiation of rat olfactory neurons counteracts apoptosis. *Cell Death Differ.* (in the press).
26. Reale, M. A. *et al.* Expression and alternative splicing of the deleted in colorectal cancer (*DCC*) gene in normal and malignant tissues. *Cancer Res.* **54**, 4493–4501 (1994).
27. Rabizadeh, S. *et al.* Mechanism of neurotrophin dependence: requirement for a novel type of domain. *Cell* (submitted).
28. Jordan, M., Schallhorn, A. & Wurm, F. M. Transfecting mammalian cells: optimization of critical parameters affecting calcium-phosphate precipitate formation. *Nucleic Acids Res.* **24**, 596–601 (1996).
29. Stennicke, H. R. & Salvesen, G. S. Biochemical characteristics of caspases-3, -6, -7, and -8. *J. Biol. Chem.* **272**, 25719–25723 (1997).

Acknowledgements. We thank B. Vogelstein and E. Fearon for the DCC cDNA, M. Tessier-Lavigne for the netrin-1 cDNA, and F. Martin for the REGb cells. This work was supported in part by grants from the NIH and the VivoRx Corporation.

Correspondence and requests for materials should be addressed to D.E.B. (e-mail: dbredesen@ljcrf.edu).

Assembly of the *Drosophila* phototransduction cascade into a signalling complex shapes elementary responses

Kristin Scott & Charles S. Zuker

Howard Hughes Medical Institute and Departments of Biology and Neurosciences, University of California at San Diego, La Jolla, California 92093-0649, USA

The subcellular compartmentalization of signalling molecules helps to ensure the selective activation of different signal-transduction cascades within a single cell¹. Although there are many examples of compartmentalized signalling molecules, there are few examples of entire signalling cascades being organized as distinct signalling complexes. In *Drosophila* photoreceptors, the InaD protein, which consists of five PDZ domains, functions as a multivalent adaptor that brings together several components of the phototransduction cascade into a macromolecular complex^{2–5}. Here we study single-photon responses in several photoreceptor mutant backgrounds, and show that the InaD macromolecular complex is the unit of signalling that underlies elementary responses. We show that the localized activity of this signalling unit promotes reliable single-photon responses as well as rapid activation and feedback regulation. Finally, we use genetic and electrophysiological tools to illustrate how the assembly of signalling molecules into a transduction complex limits signal amplification *in vivo*.

Phototransduction in *Drosophila* occurs through a phosphoinositide-mediated and calcium-regulated G-protein-coupled transduction pathway, in which light-mediated activation of rhodopsin leads to the sequential activation of a heterotrimeric G protein, an eye-specific phospholipase C (PLC) and the gating of the transient receptor potential (TRP) and TRPL light-activated ion channels^{6,7}. The InaD scaffold protein assembles several components of this cascade, including TRP, PLC and protein kinase C (PKC), into an organized protein–protein complex^{2–5}. Null *inaD* mutants have a marked disruption of the subcellular distribution of signalling molecules, a loss of transduction complexes and severely impaired photoresponses².

Drosophila photoreceptors can 'report' activity with high sensitivity and specificity: photoreceptor cells are sensitive to single photons, and the signalling pathway can be turned on and off with millisecond kinetics (phototransduction in *Drosophila* is the fastest known G-protein cascade, taking just a few tens of milliseconds to go from light activation of rhodopsin to the generation of a receptor potential)^{6,7}. In photoreceptor neurons, the elementary response to a single photon of light is known as a quantum

bump^{8–10}. A quantum bump results from the opening (or closing) of light-activated ion channels in response to one activated rhodopsin molecule, and reflects the amplification of the entire visual cascade. We postulated that the signalling kinetics of photoreceptor neurons may be a direct consequence of organizing signalling molecules into architecturally defined signalling units. We reasoned that InaD may function as a master scaffold to assemble the phototransduction machinery^{4,5}; and therefore sought to determine whether individual InaD macromolecular complexes embody elementary responses.

We studied light-induced currents in wild-type and mutant photoreceptor neurons by performing whole-cell patch-clamp recordings under conditions that activate single rhodopsin molecules¹¹. Mutants lacking InaD (*inaD*¹ mutants) have single-photon responses that are grossly disorganized, with large defects in amplitude, latency and deactivation (Fig. 1 and Table 1). Quantum bumps in *inaD*¹ mutants have amplitudes that are less than one-fifth those of controls, latencies that are roughly six times greater than wild-type latencies and responses that fail to terminate normally, with single photons producing several microbumps. These results demonstrate a critical role for InaD in quantum-bump generation and substantiate a requirement for the InaD signalling complex in both activation and feedback regulation.

If the InaD signalling complex provides the molecular framework of a quantum bump, it should be possible to manipulate quantum bumps by selectively manipulating the InaD complex. We therefore studied quantum bumps in *inaD*² photoreceptors, a mutant defective in the fifth PDZ (postsynaptic density protein, discs-large, ZO-1) domain of InaD; this mutant fails to recruit PLC to the transduction complex². As predicted, *inaD*² mutants show severe bump defects, with latency times over eight times larger than those of wild-type controls (Fig. 1 and Table 1). This activation phenotype is consistent with a dramatic reduction of PLC levels in transduction complexes, and resembles the phenotype produced by *norpA* hypomorphic alleles (see below). However, *inaD*² mutants, unlike *norpA* hypomorphs, have nearly normal levels of PLC², showing that it is not simply the presence of PLC that is important for function, but rather its location.

A quantum bump represents the opening of many ion channels¹²; therefore, an InaD signalling complex (transducisome) must contain several InaD molecules and ion channels assembled into a macromolecular complex. If a quantum bump represents the output of a localized signalling complex, then fundamental properties of a quantum bump, such as amplitude, latency, rise and decay¹³, should reflect the activity of individual signalling complexes. A prominent feature of invertebrate quantum bumps is their significant variability in amplitude (Fig. 1). This broad distribution in amplitude has been thought to result from randomness in the amplification process, leading to stochastic fluctuations in the number of activated ion channels per rhodopsin molecule¹³. We presumed that the variability in bump amplitude may instead reflect variation in the composition of different signalling com-

Table 1 Kinetics of quantum bumps

Mutant	Number of bumps	Number of trials	Mean frequency of events	Relative light stimulus	Latency (ms)	Amplitude (pA)
WT	147	480	0.36	1	66 ± 24	36 ± 19
<i>inaD</i> ¹	73	336	0.24	1	385 ± 374	8 ± 7
<i>inaD</i> ²	68	240	0.33	1	481 ± 389	19 ± 11
<i>chp</i>	65	192	0.41	1	160 ± 120	27 ± 16
<i>norpA</i> ^{H52}	87	360	0.28	1	372 ± 270	29 ± 17
<i>norpA</i> ^{P16}	274	ND	ND	1	~60,000	27 ± 13
WT pupae	139	239	0.87	1	62 ± 52	18 ± 23
<i>Gαq</i> ¹ pupae	118	220	0.77	1,000	104 ± 96	18 ± 20

The table shows mean values (±s.d.) for latency, amplitude and relative light stimulus (defined as the amount of light required to generate a quantum bump normalized to wild-type stimulation) in several mutant backgrounds. Bumps were evoked by light flashes of 10 ms ($\log[I] = -6.5$ for wild-type; this was set as 1 in the relative light calculations); only some of the flashes (or trials) generated a bump. The mean frequency of events was calculated by assuming a Poisson distribution in which the probability of zero bumps (P_0) = $e^{-m} = (1 - \text{no. of bumps/no. of trials})$, and solving for m (ref. 8). The data for the *Gαq*¹ mutant were obtained from late-stage pupae; therefore, wild-type controls from the same stage are also included (*Gαq*¹ values were calculated from ref. 11).

chp, *chaoptin*; ND, not determined; WT, wild-type.

plexes (that is, variation in the number of channels contained in a complex). As a photoreceptor cell contains 10^8 molecules of rhodopsin¹⁴, the probability of generating multiple single-photon responses from the same rhodopsin molecule is infinitesimal, and so the variance could result from activity of different signalling complexes. A prediction of this postulate is that if it were possible to stimulate the same rhodopsin molecule over and over (that is, the

same signalling complex repeatedly), then quantum-bump amplitudes should be highly reproducible with only a small variance. This experiment could be done either by engineering flies that have one or a few rhodopsin molecules per cell, or by generating modified photoreceptors in which one rhodopsin molecule signals continuously after activation by a single photon of light. We showed previously that calmodulin is required for rhodopsin shut-off by

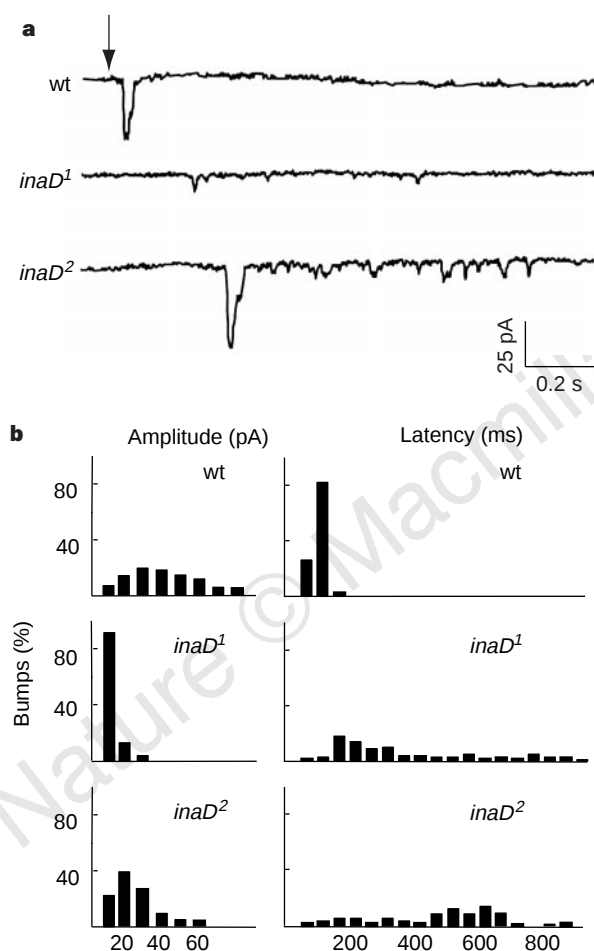


Figure 1 Disrupting the InaD signalling complex disrupts quantum bumps. **a**, Single-photon responses in *inaD* mutants. Cells were stimulated with a 10-ms flash of light of wavelength 580 nm ($\log[I] = -6.5$; see Methods) at the time indicated by the arrow. The *inaD¹* null mutant shows defects in amplitude, latency and deactivation of quantum bumps, consistent with a loss of signalling complexes. The *inaD²* allele contains a mutation in its PLC-interaction site, and exhibits disruption of latency and deactivation. wt, wild-type. **b**, Amplitude and latency distributions of *inaD* quantum bumps. Responses were measured for 73 *inaD¹* bumps (from 6 cells) and 67 *inaD²* bumps (from 5 cells). The probability of a bump occurring was 0.22 for *inaD¹* and 0.28 for *inaD²* mutants (assuming a Poisson distribution of events and solving for P values = $1 - (\text{trials with zero quantum bumps}/\text{total trials})^8$). Because *inaD* mutants produce more than one response per flash, only the amplitude and latency of the first response were measured. Table 1 shows the quantification of response parameters.

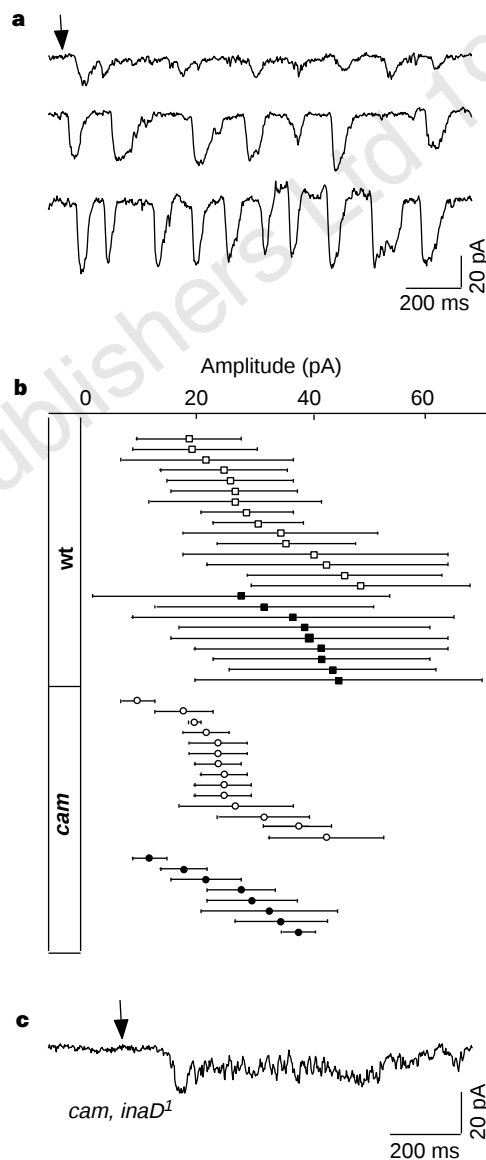


Figure 2 A quantum bump is an organized unit of signalling. **a**, Single-photon responses from a *cam* mutant photoreceptor. Each trace represents quantum bumps produced by a single activated photoreceptor; the three different traces represent the activity produced by three different rhodopsin molecules. Bath solution contained 0.4 mM CaCl_2 . **b**, The spread of quantum bumps produced by different activated receptors compared with the spread of bumps produced by one activated receptor. Responses of 24 wild-type (wt) cells and 22 *cam* mutants are shown. Each trace represents the mean $\pm$ s.d. of seven bumps generated by consecutive flashes of light (wt), or the mean $\pm$ s.d. of seven bumps generated by a single photon of light (*cam*). Cells were stimulated with a 10-ms flash of 580-nm light, $\log[I] = -6.5$. The spread of all of the different *cam* responses mimics the variability seen in normal single-photon responses. The open symbols indicate that the bath solution contained 1.5 mM CaCl_2 , and the filled symbols indicate that the bath solution contained 0.4 mM CaCl_2 . **c**, Quantum bumps from *cam, inaD¹* double mutants show that loss of InaD disrupts localized signalling.

regulating arrestin function, and that calmodulin (*cam*) mutants produce many quantum bumps in response to a single photon of light¹⁵. Therefore, we used a hypomorphic allele of *cam* to analyse quantum-bump variance following activation of discrete transducisomes. Figure 2a shows sample traces of trains of quantum bumps from one *cam* mutant photoreceptor, illustrating both the high reproducibility of bump amplitudes following activation of the same rhodopsin molecule and the wide variability produced by the activation of different rhodopsin molecules. These results indicate that differences in bump amplitude do not simply stem from stochastic fluctuations in the number of activated channels. Instead, because a given receptor can consistently activate a similar number of ion channels (Fig. 2b), these results indicate that a quantum bump reflects the activation of a reliable unit of signalling. If the InaD transducisome represents this signalling unit, then *cam* mutants lacking InaD should no longer exhibit discrete trains of quantum bumps. As predicted, *cam, inaD*¹ double mutants generate a noisy, continuous response to a single photon of light, and do not produce individual bumps (Fig. 2c).

As *Drosophila* phototransduction takes place in a microvillar organelle (each photoreceptor cell contains ~60,000 microvilli, and

each microvillus has ~1,000 rhodopsins)⁶, we wondered whether the anatomical structure of a microvillus places physical boundaries on the organization of signalling complexes. Thus, we studied quantal responses in mutant photoreceptors that lack organized microvilli (*chaoptin* mutants)^{16,17}. Quantum-bump amplitudes in *chaoptin* mutants are not significantly different from wild-type amplitudes (Table 1), showing that the microvillar architecture does not contribute to the structural organization underlying a quantum bump.

It is thought that the primary function of signalling cascades is to promote amplification. However, the assembly of signalling molecules into a protein–protein macromolecular complex would be expected to severely limit amplification, as the number of downstream elements in the cascade would be physically limited. Because phototransduction takes place in InaD complexes, we wondered whether there is any amplification upstream of the light-activated ion channels. To investigate this issue *in vivo*, we used the following logic: if a rhodopsin molecule activates many G proteins, then a reduction in the cellular levels of the G protein should result in fewer channels opening and a corresponding reduction in bump amplitude (as is seen in mutants that contain fewer light-activated channels¹²). However, if there is no amplification between rhodopsin and the G protein, then reducing the levels of the G protein should not affect bump size but should instead affect the frequency or latency of the response. Similarly, if there is no amplification at the G-protein–PLC interface, reducing the levels of PLC should not affect bump amplitude.

We used genetically engineered flies that express low levels of G protein (~1% of normal levels; see ref. 11) and severe hypomorphic alleles of *norpA* (which produce levels of PLC that are ~10% of wild-type levels)^{18,19} to reduce the levels of these signalling molecules *in vivo*. Figure 3 and Table 1 show the distributions of quantum-bump amplitudes, latencies and relative bump frequencies for wild-type, *Gaq*¹ and *norpA* photoreceptors, demonstrating that a marked reduction in the levels of the G protein¹¹ or PLC does not significantly affect the size or shape of a quantum bump, although latency or frequency are significantly altered. These results demonstrate that there is little amplification upstream of second-messenger production, and suggest that one rhodopsin activates one G protein which in turn activates one PLC, leading to the opening of many ion channels. Interestingly, the G-protein hypomorphs exhibit nearly normal latencies but a 1,000-fold reduction in sensitivity (that is, ~1,000 rhodopsin molecules must be activated to generate a bump), whereas the most severe PLC hypomorphs exhibit a ~1,000-fold increase in latency but normal sensitivity. Why does a reduction in G protein levels mainly affect frequency, whereas a reduction in PLC levels affects latency? These findings can be explained easily by assuming that the lifetime of an activated rhodopsin is very short (being limited by arrestin-dependent shut-off) and so it is necessary to excite many rhodopsins before activating one in close proximity to a G protein. In contrast, the lifetime of an activated G protein is very long in the absence of PLC, possibly because PLC acts as a GTPase-activating protein to deactivate G_q, and so a G protein stays active until it encounters a PLC. However, once a rhodopsin encounters a G protein, or once a G protein finds a PLC, the generation of a bump occurs normally.

Collectively, our results illustrate how the organization of signalling molecules into discrete signalling units provides fundamental advantages, such as rapid responses and feedback regulation, to a transduction cascade. Our study of single-photon responses and their relationship to the InaD macromolecular complex also uncovered several unexpected findings. First, we showed that a physiologically described signalling event, such as a quantum bump, has a well-defined molecular framework, and that the organization of signalling molecules into a macromolecular complex ensures stable and reliable unitary responses. Second, we demonstrated that this organization limits the amplification of the cascade *in vivo*. Why

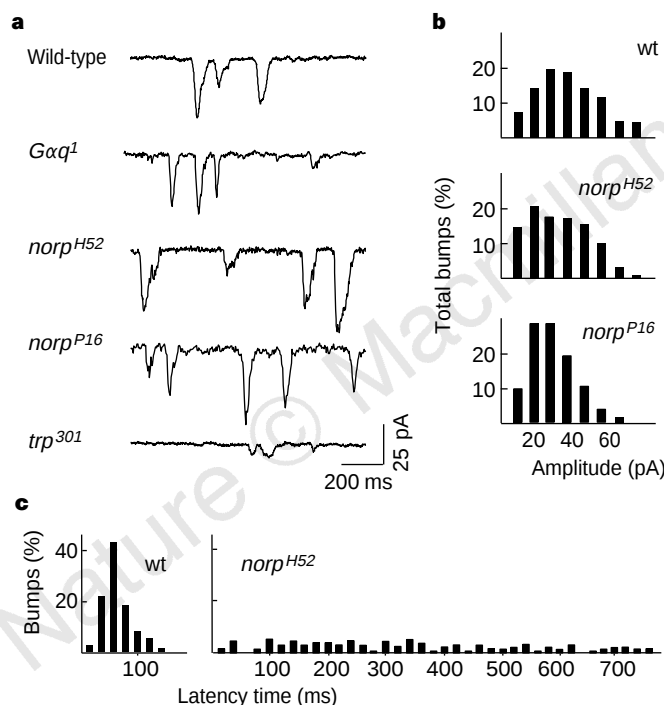


Figure 3 Amplification occurs downstream of PLC activation. **a**, Quantum bumps from a wild-type photoreceptor, from a photoreceptor containing ~1% of the normal levels of the eye-specific G α subunit (*Gaq*¹), and from photoreceptors that express reduced levels of PLC (*norpA*^{H52} and *norpA*^{P16}) are shown. Quantum bumps from a photoreceptor with low levels of the light-activated ion channel TRP (*trp*³⁰¹) are also shown; this photoreceptor exhibits significant reduction in bump size¹². These results indicate that signal amplification occurs at the terminal signalling events, downstream of PLC activation. Stimulation was with continuous light of log[I] = -7 for wild-type, *norpA*^{H52} and *trp*³⁰¹, and log[I] = -4 for *Gaq*¹. **b**, Amplitude distributions of quantum bumps in the *norpA* hypomorphs. For studies with wild-type and *norpA*^{H52} alleles, cells were stimulated with a 10-ms flash of 580-nm light of log[I] = -6.5. The probability of a bump occurring (*P*) was 0.30 for wild-type and 0.28 for *norpA*^{H52} photoreceptors. **c**, Distributions of latency times for the *norpA*^{H52} hypomorph. Latency was measured as the time from the onset of the light stimulus (10 ms, log[I] = -6.5) to the peak amplitude of the response. For the *norpA*^{P16} allele, the mean latency time was estimated as ~60 s, derived from the exponential decay of the bump rate following stimulation with a 10-ms flash of light of log[I] = -4.5. See Table 1 for quantification of response parameters.

have multiple components in this transduction pathway if there is little amplification? Having many steps in the same signalling cascade allows fine control by providing several points of regulation. In the case of phototransduction, this may allow for regulatory mechanisms that endow photoreceptor cells with high temporal resolution, refined adaptation and a broad dynamic range of response^{20,21}. □

Methods

We isolated photoreceptors from adult flies (>6 h after eclosion) and performed whole-cell, voltage-clamp recordings. The bath solution contained 124 mM NaCl, 4 mM KCl, 10 mM HEPES, 5 mM proline and 25 mM sucrose, pH 7.1. Unless otherwise stated, the bath solution also contained 1.5 mM CaCl₂. The pipette solution contained 95 mM potassium gluconate, 40 mM KCl, 10 mM HEPES, 2 mM MgCl₂ and 0.2 mM EGTA, pH 7.15.

Photoreceptors were clamped at a holding potential of -70 mV. Signals were sampled at 1 kHz, then filtered at 500 Hz. In all experiments, light was filtered through a bandpass filter ($\lambda = 580 \pm 10$ nm) and neutral-density filters, and focused onto the photoreceptor cells through a 0.5 numerical aperture, $\times 40$ objective. Photoreceptors were stimulated with a 75 W xenon source; I_0 is the maximum intensity at 580 nm produced by the light source (0.04 mW cm^{-2}).

Single-photon responses were generated and analysed as described^{8,11}. For analysis of quantum bumps in the temperature-sensitive *norPA*^{H52} allele, flies were heat-shocked for 1 h at 37 °C and set at 22 °C for 30 min; patch-clamp analysis was done as usual. For *norPA*^{P16}, bumps were continuously produced immediately after break-in (probably induced by the microscope light required for patching), with the bump rate declining over a period of several hundred seconds. Amplitudes were measured when the bump rate declined to <1 bump per second.

Received 7 August; accepted 18 September 1998.

1. Tsunoda, S., Sieralta, J. & Zuker, C. Specificity in signaling pathways: assembly into multimolecular signaling complexes. *Curr. Opin. Genet. Dev.* **8**, 419–422 (1998).
2. Tsunoda, S. *et al.* A multivalent PDZ-domain protein assembles signalling complexes in a G-protein-coupled cascade. *Nature* **388**, 243–249 (1997).
3. Chevesich, J., Kreuz, A. J. & Montell, C. Requirement for the PDZ domain protein, INAD, for localization of the TRP store-operated channel to a signaling complex. *Neuron* **18**, 95–105 (1997).
4. Shieh, B. H. & Zhu, M. Y. Regulation of the TRP Ca²⁺ channel by INAD in *Drosophila* photoreceptors. *Neuron* **16**, 991–998 (1996).
5. Huber, A. *et al.* The transient receptor potential protein (Trp), a putative store-operated Ca²⁺ channel essential for phosphoinositide-mediated photoreception, forms a signaling complex with NorpA, InaC and InaD. *EMBO J.* **15**, 7036–7045 (1996).
6. Ranganathan, R., Malicki, D. M. & Zuker, C. S. Signal transduction in *Drosophila* photoreceptors. *Annu. Rev. Neurosci.* **18**, 283–317 (1995).
7. Zuker, C. S. The biology of vision in *Drosophila*. *Proc. Natl Acad. Sci. USA* **93**, 571–576 (1996).
8. Baylor, D. A., Lamb, T. D. & Yau, K.-W. Responses of retinal rods to single photons. *J. Physiol. (Lond.)* **288**, 613–634 (1979).
9. Yeandle, S. *Studies on the slow potential and the effect of cations on the electrical responses of Limulus ommatidium*. Thesis, Johns Hopkins Univ. (1957).
10. Fortes, M. T. F. & Hodgkin, A. L. Changes in time scale and sensitivity in the ommatidia of *Limulus*. *J. Physiol. (Lond.)* **172**, 239–263 (1964).
11. Scott, K., Leslie, A., Sun, Y., Hardy, R. & Zuker, C. Gαq protein function *in vivo*: genetic dissection of its role in photoreceptor cell physiology. *Neuron* **15**, 919–927 (1995).
12. Niemeyer, B. A., Suzuki, E., Scott, K., Jalink, K. & Zuker, C. S. The *Drosophila* light-activated conductance is composed of the two channels TRP and TRPL. *Cell* **85**, 651–659 (1996).
13. Stieve, H. (ed.) *Bumps, the Elementary Excitatory Responses of Invertebrates* (Dahlem Konferenzen, Springer, Berlin, 1986).
14. Johnson, E. & Pak, W. Electrophysiological study of *Drosophila* rhodopsin mutants. *J. Gen. Physiol.* **88**, 651–673 (1986).
15. Scott, K., Sun, Y. M., Beckingham, K. & Zuker, C. S. Calmodulin regulation of *Drosophila* light-activated channels and receptor function mediates termination of the light response *in vivo*. *Cell* **91**, 375–383 (1997).
16. Reinke, R., Krantz, D. E., Yen, D. & Zipursky, S. L. Choptin, a cell surface glycoprotein required for *Drosophila* photoreceptor cell morphogenesis, contains a repeat motif found in yeast and human. *Cell* **52**, 291–301 (1988).
17. van Vactor, D. Jr, Krantz, D. E., Reinke, R. & Zipursky, S. L. Analysis of mutants in Choptin, a photoreceptor cell-specific glycoprotein in *Drosophila*, reveals its role in cellular morphogenesis. *Cell* **52**, 281–290 (1988).
18. Pearn, M. T., Randall, L. L., Shortridge, R. D., Burg, M. G. & Pak, W. L. Molecular, biochemical, and electrophysiological characterization of *Drosophila* *norPA* mutants. *J. Biol. Chem.* **271**, 4937–4945 (1996).
19. Deland, M. C. & Pak, W. L. Reversibly temperature sensitive phototransduction mutant of *Drosophila melanogaster*. *Nature* **244**, 84–86 (1973).
20. Scott, K. & Zuker, C. Lights out: deactivation of the phototransduction cascade. *Trends Biochem. Sci.* **22**, 350–354 (1997).
21. Hardie, R. C. & Minke, B. Phosphoinositide-mediated phototransduction in *Drosophila* photoreceptors: the role of Ca²⁺ and *trp*. *Cell Calcium* **18**, 256–274 (1995c).

Acknowledgements. We thank members of the Zuker lab for reading the manuscript and for discussions. C.S.Z. is an investigator of the Howard Hughes Medical Institute. This work was funded in part by a grant from the National Eye Institute.

Correspondence and requests for materials should be addressed to C.S.Z. (e-mail: czuker@flyeye.ucsd.edu).

The G protein Gα12 stimulates Bruton's tyrosine kinase and a rasGAP through a conserved PH/BM domain

Yun Jiang*, Wei Ma*, Yong Wan*, Tohru Kozasa†, Seisuke Hattori‡, & Xin-Yun Huang*

* Department of Physiology, Cornell University Medical College, 1300 York Avenue, New York, New York 10021, USA

† Department of Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas 75235, USA

‡ Division of Biochemistry and Cellular Biology, National Institute of Neuroscience, Kodaira, Tokyo 187, Japan

Heterotrimeric guanine-nucleotide-binding proteins (G proteins) are signal transducers that relay messages from many receptors on the cell surface to modulate various cellular processes^{1–4}. The direct downstream effectors of G proteins consist of the signalling molecules that are activated by their physical interactions with a Gα or Gβγ subunit. Effectors that interact directly with Gα12 G proteins have yet to be identified^{5,6}. Here we show that Gα12 binds directly to, and stimulates the activity of, Bruton's tyrosine kinase (Btk) and a Ras GTPase-activating protein, Gap1^m, *in vitro* and *in vivo*. Gα12 interacts with a conserved domain, composed of the pleckstrin-homology domain and the adjacent Btk motif, that is present in both Btk and Gap1^m. Our results are, to our knowledge, the first to identify direct effectors for Gα12 and to show that

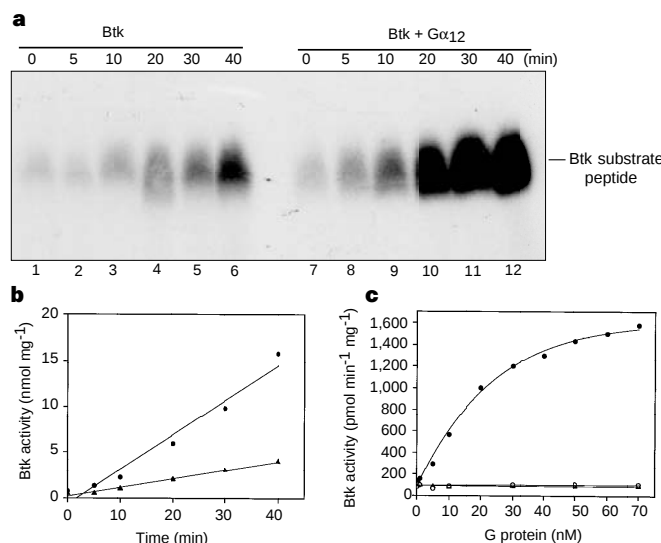


Figure 1 Stimulation of Btk kinase activity by Gα12. **a**, **b**, Time course of Btk kinase activity. **a**, Btk was incubated with substrate peptide for various times in the absence (lanes 1–6) or presence (lanes 7–12) of 5 nM Gα12-GTPγS. **b**, After longer exposure, the corresponding bands on the gel were cut out and phosphorylation was quantified by scintillation counting. Btk activity was expressed as nmoles PO₄ incorporated into peptide substrate per mg Btk. Data shown are representative of three similar experiments. Circles, activity in the presence of Gα12-GTPγS; triangles, activity in the absence of Gα12. **c**, Stimulation of Btk kinase activity by different concentrations of Gα12-GTPγS. The indicated concentrations of α-subunits were reconstituted with 1 ng Btk for 20 min at 30 °C. Increasing amounts of Gα12-GTPγS (filled circles) increased Btk activity. Gα12-GTPγS (open circles) or Gα12-GDP (filled triangles, barely visible) had no effect on Btk activity. Data are representative of 4–6 experiments.

there is a direct link between heterotrimeric and monomeric G proteins.

We have shown previously that the α -subunit of the G_q class of G proteins ($G_{\alpha q}$) can bind directly to Btk and activate its kinase activity⁷. While testing the effects of other families of G proteins on Btk activity, we found that purified α -subunits of the G_i and G_s families had no effect, but purified $G_{\alpha 12}$ directly stimulated the tyrosine kinase activity of Btk (Fig. 1). Marked stimulation of Btk activity by purified $G_{\alpha 12}$ required preactivation with GTP γ S (Fig. 1). Reconstitution of Btk with purified $G_{\alpha 12}$ -GDP did not alter Btk activity (Fig. 1). Therefore, $G_{\alpha 12}$ directly stimulates Btk's kinase activity *in vitro*.

To understand the mechanism of activation of Btk we determined the location of the $G_{\alpha 12}$ -binding site on Btk (Fig. 2). The primary structure of Btk is organized into the following series of domains: a pleckstrin-homology (PH) domain (residues 1–138), a Tec-homology (TH) domain (139–215), a Src-homology (SH) 3 domain (216–280), an SH2 domain (281–377), and a kinase domain (378–659)⁸ (Fig. 2a). The amino-terminal portion of the TH domain contains a unique Btk motif (BM, residues 139–174) which is common to the Btk-family tyrosine kinases⁹, and its carboxy-terminal portion contains a proline-rich region (residues 175–215). We produced fusion proteins containing glutathione S-transferase (GST) and one of the domains above (GST-PH, GST-TH, GST-SH3, GST-SH2 and GST-KIN) and tested their ability to bind to $G_{\alpha 12}$ (Fig. 2b, c). None of these individual, separated domains bound $G_{\alpha 12}$ from HEK293 cells transfected with activated $G_{\alpha 12}$ ($G_{\alpha 12}$ (Q229L)) (Fig. 2c). We next made GST-fusion proteins containing two neighbouring domains (GST-PHSH, GST-THSH3, and GST-SH3SH2; Fig. 2b). We found that only GST-PHSH could bind $G_{\alpha 12}$ (Fig. 2c). More specifically, sequences composed of the C-terminal portion of the PH domain (PH-C) and the Btk motif are sufficient for $G_{\alpha 12}$ binding (construct GST-PH-C/BM) (Fig. 2c). Thus, $G_{\alpha 12}$ binds to Btk through a region spanning the PH domain and the Btk motif.

To confirm that the PHTH fragment of Btk interacted directly with $G_{\alpha 12}$, we tested binding using purified components *in vitro*. Purified $G_{\alpha 12}$ -GTP γ S bound purified GST-PHSH fusion protein easily, whereas $G_{\alpha 12}$ -GDP did not (Fig. 3a). There was no detectable binding of $G_{\alpha 12}$ -GTP γ S to control GST or GST-SH3SH2 fusion proteins. We also determined whether $G_{\alpha 12}$ interacts with

Btk *in vivo* by using a co-immunoprecipitation assay. Endogenous Btk from lymphoma DT40 cells transfected with activated $G_{\alpha 12}$ (Q229L) was immunoprecipitated by antibodies directed against the C terminus of Btk. Immunoblot analysis showed that $G_{\alpha 12}$ was found in the immunoprecipitate obtained with anti-Btk antibody, but not in the immunoprecipitate obtained with control antiserum (Fig. 3b). Conversely, an anti- $G_{\alpha 12}$ antibody could co-immunoprecipitate Btk (Fig. 3c). These results show that $G_{\alpha 12}$ interacts with Btk *in vivo*. $G_{\alpha 12}$ can also stimulate endogenous Btk activity *in vivo* in DT40 lymphoma cells. Overexpression of $G_{\alpha 12}$ (Q229L) in DT40 lymphoma cells led to increased (roughly three times control levels) kinase activity of endogenous Btk (Fig. 3d, e). Thus, $G_{\alpha 12}$ can stimulate Btk activity *in vitro* and *in vivo*.

A database search showed that the tandem PH/BM module (a PH domain followed immediately by the Btk motif) is also present in the recently identified Ras GTPase-activating-protein (rasGAP) Gap1^m (Fig. 4a)^{10,11}. Gap1^m is a member of the Gap1 family of rasGAPs. Other members of this family, including *Drosophila* Gap1 (ref. 12), the inositol-1,3,4,5-tetrakisphosphate (InsP₄) receptor Gap1^{IP4BP} (ref. 13), and the R-Ras GAP¹⁴, all contain the conserved tandem PH/BM module. To determine whether $G_{\alpha 12}$ can stimulate

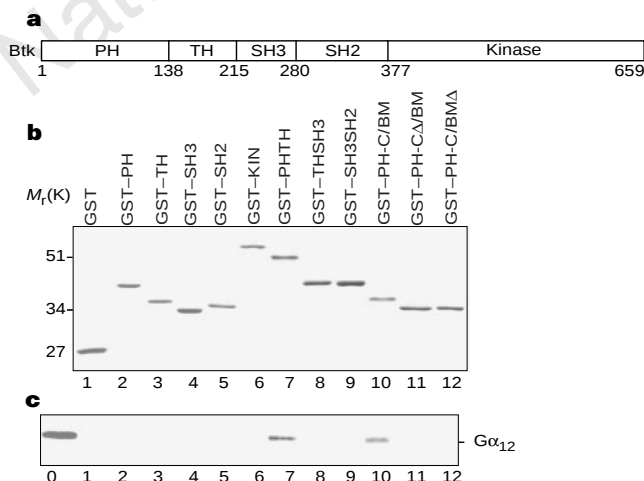


Figure 2 $G_{\alpha 12}$ association with Btk *in vitro*. **a**, Representation of Btk and its structural domains. The numbers at the bottom indicate positions of amino-acid residues. **b**, Coomassie blue staining of SDS-PAGE of purified GST-fusion proteins. **c**, Binding of $G_{\alpha 12}$ to the PHTH region of Btk. Lane 0 was loaded with 40 μ l whole-cell extract as a control. Lanes 1–12 were loaded with GST-fusion proteins as indicated in **b**.

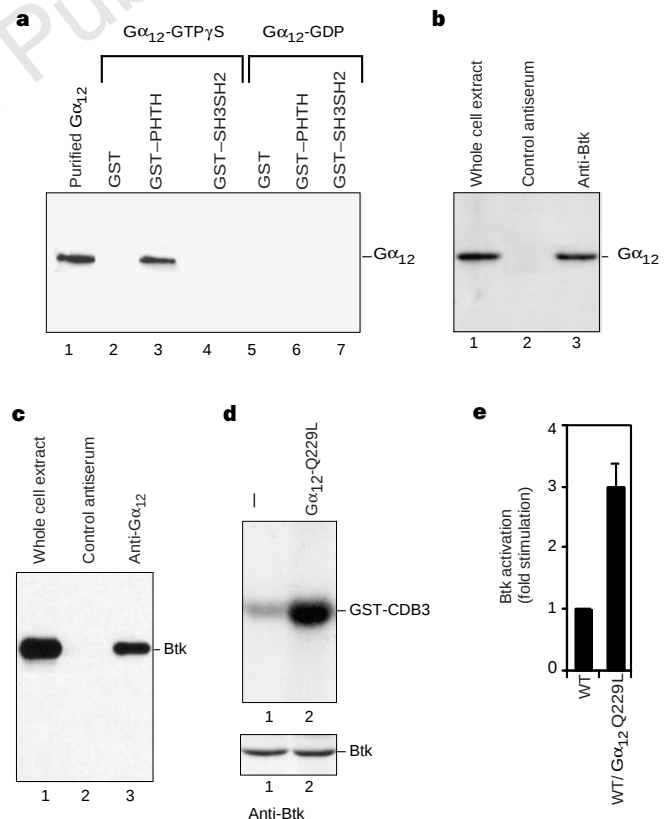


Figure 3 Association of $G_{\alpha 12}$ with Btk *in vivo* and *in vitro*. **a**, Direct interaction of purified $G_{\alpha 12}$ -GTP γ S with purified GST-PHSH fusion protein *in vitro*. Lane 1 loaded with purified $G_{\alpha 12}$ as control. **b**, **c**, $G_{\alpha 12}$ interacts with Btk *in vivo*. Whole-cell lysates were subjected to immunoprecipitation with **b**, anti-Btk antibody or **c**, anti- $G_{\alpha 12}$ antibody, and immunoprecipitated proteins were analysed by immunoblotting with **b**, anti- $G_{\alpha 12}$ antibody or **c**, anti-Btk antibody. Precipitation with anti-MEK antibody (lane 2 in **b**, **c**) was included as control antiserum. Also, lane 1 (in **b**, **c**) was loaded with whole-cell extract as a control. **d**, Stimulation of Btk kinase activity by $G_{\alpha 12}$ *in vivo*. The ectopic expression of the constitutively active $G_{\alpha 12}$ mutant $G_{\alpha 12}$ (Q229L) stimulates Btk kinase activity towards GST-CDB3. Bottom, western blotting with an anti-Btk antibody showed that the amount of Btk loaded in each lane was comparable. **e**, Quantification of Btk kinase stimulation. Data represent mean $\pm$ s.d. of three experiments.

Gap1^m activity directly, we purified recombinant Gap1^m from *Escherichia coli*. The activity of Gap1^m was monitored by its ability to stimulate the hydrolysis of GTP bound to recombinant Ha-Ras, also purified from *E. coli*. All proteins were quite pure, as judged by silver or Coomassie brilliant blue stainings of SDS gels and by western blot analysis with specific antibodies (Fig. 4b). [γ -³²P]GTP was bound to Ha-Ras, and GTP hydrolysis by Ha-Ras was measured by ³²P_i release in the absence or presence of Gap1^m protein and in the absence or presence of G α 12 (Fig. 4c, d). As we measured only one round of hydrolysis of the prebound [γ -³²P]GTP we used a quite high concentration of Ha-Ras. The reaction mixture contained 750 nM Ha-Ras¹³. In control reactions, the GTPase activity of Ha-Ras was ~40 pmol P_i released per mg Ha-Ras. In the presence of Gap1^m, GTP hydrolysis was increased by roughly eightfold (Fig. 4c). Although G α 12 alone had little effect on the GTPase activity of Ha-Ras, G α 12 markedly increased the Ha-Ras GTPase activity ~30-fold in the presence of Gap1^m, indicating that G α 12 can enhance Gap1^m activity towards Ha-Ras (Fig. 4c).

We tested the functional specificity of the interaction of Gap1^m with different families of G α proteins. The effects of G α _s, G α _{i1} and G α _q and G α 12 on Gap1^m-mediated stimulation of Ha-Ras GTPase activity are summarized in Fig. 4d. None of G α _s, G α _{i1}, G α _q and G β γ ₂ had any effect on Gap1^m stimulation of Ha-Ras GTPase

activity (Fig. 4d). Therefore, G α 12 stimulates Gap1^m activity specifically. Although the concentration of G α 12-GTP γ S required to achieve maximal stimulation of Btk is slightly higher than that required for the stimulation of Gap1^m, the two assay conditions (such as reaction time and quantification methods) are different, making it hard to compare their efficacies directly. Nevertheless, the concentrations of G α 12-GTP γ S required to stimulate both effectors are in the lower nanomolar range. For example, 5 nM G α 12-GTP γ S increased both the Btk kinase activity and the activity of Gap1^m by roughly three- to fourfold.

To confirm that the PH/BM region of Gap1^m is indeed responsible for binding to G α 12, as we expected, we determined the location of the G α 12-binding site on Gap1^m (Fig. 5). Gap1^m contains two domains that are homologous to the C2 regulatory region of protein kinase C (C2A and C2B, residues 1–294), a GAP-related domain (GRD, 295–602), a PH domain (603–703), an immediately adjacent BM (704–739) and the C-terminal tail (740–847)¹¹ (Fig. 4a). We produced GST-fusion proteins containing the full-length Gap1^m (GST-Gap1^m), the two C2 domains and the GRD (GST-C2GRD), the GRD, and PH and BM domains (GST-GRDPHBM), the C2A and C2B domains (GST-C2), the GRD domain (GST-GRD), or the PH and BM domains (GST-PHBM) (Fig. 5a). We analysed the ability of these fusion proteins to bind

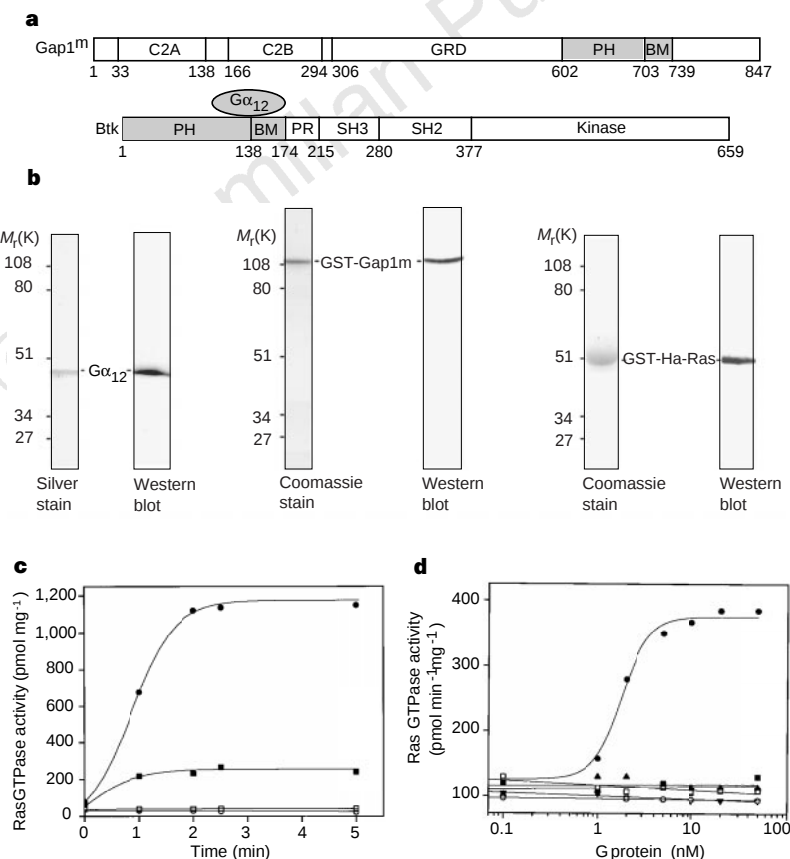


Figure 4 Stimulation of Gap1^m activity by G α 12. **a**, Representation of the structural domains of Gap1^m and Btk, including the G α 12-binding domain of Btk. The numbers at the bottom indicate positions of amino-acid residues. PR, proline-rich domain. **b**, Purification of G α 12, Gap1^m, and Ha-Ras. Left, silver staining of purified G α 12 from recombinant baculovirus-infected Sf9 cells; western blotting confirmed the identity of G α 12. Centre and right, Coomassie staining of purified GST-Gap1^m or GST-Ha-Ras from *E. coli*; the identity of Gap1^m and Ha-Ras was confirmed by western blotting with anti-Gap1^m or anti-Ha-Ras antibody. Prestained relative molecular mass (*M_r*) standards (BioRad) are shown on the left. **c**, Time course of Ha-Ras GTPase activity. 0.35 μ g [γ -³²P]GTP-bound Ha-Ras was incubated for various times in the absence (open circles) or presence (open

squares) of 10 nM G α 12-GTP γ S, or with 2.6 ng Gap1^m (filled squares) or 10 nM G α 12-GTP γ S plus 2.6 ng Gap1^m (filled circles). Ha-Ras GTPase activity was expressed as pmol PO₄ released per mg Ha-Ras. Data shown are representative of four similar experiments. **d**, Stimulation of Gap1^m activity by different concentrations of G α 12-GTP γ S. The indicated concentrations of G-protein subunits were reconstituted with Ha-Ras and Gap1^m for 3 min at 25 °C. Increasing amounts of G α 12-GTP γ S (filled circles) increased Gap1^m-mediated stimulation of Ha-Ras GTPase activity. G α 12-GDP (filled upward-pointing triangles), or G α s-GTP γ S (open circles) had no effect on Gap1^m activity. Data are representative of four experiments.

G α 12. The PH/BM fragment is sufficient for G α 12 binding (Fig. 5b). Fusion proteins containing the PH/BM domains bound G α 12 from HEK293 cells transfected with G α 12(Q229L). Other fusion proteins that lacked the PH/BM domains did not bind G α 12 (Fig. 5b). Thus, G α 12 indeed binds to the PH/BM module of Gap1^m.

Purified GTP γ S-bound G α 12 interacts directly and specifically with the purified GST-PHBM module of Gap1^m *in vitro* (Fig. 6a). Co-immunoprecipitation experiments with G α 12(Q229L)- and Gap1^m-transfected cells showed that G α 12 was found in the immunoprecipitate obtained using an anti-Gap1^m antibody (Fig. 6b). Conversely, an anti-G α 12 antibody co-immunoprecipitated Gap1^m (Fig. 6c). Thus G α 12 can interact with Gap1^m *in vivo* and *in vitro*.

As Gap1^m can accelerate the GTPase activity of Ras, to reduce the amount of GTP-bound Ras, and as G α 12 can stimulate Gap1^m activity, we concluded that, in Gap1^m-expressing cells, G α 12 should be able to reduce the level of Ras-GTP. In COS-7 cells, activated G α 12 inhibits epidermal growth factor (EGF)-mediated stimulation of mitogen-activated protein kinase (MAPK)¹⁵. As Ras mediates the activation of MAPK by EGF, we tested whether G α 12 could inhibit the stimulation of Ras by EGF in COS-7 cells. Gap1^m is expressed in COS-7 cells (Fig. 6d). Stimulation of COS-7 cells with EGF led to the increase of Ras-GTP levels (Fig. 6e). We detected the amount of Ras-GTP with the activated Ras-interaction assay¹⁶, using the Ras-binding domain of Raf-1 in a GST-fusion protein to precipitate Ras-GTP from cell lysates. Precipitated Ras-GTP was detected by western blotting with anti-Ras antibodies. Transfection of G α 12(Q229L) reduced the stimulation of Ras by EGF (Fig. 6e). Thus, G α 12 could negatively regulate Ras signalling in COS-7 cells.

To further investigate the physiological relevance of the interactions between G α 12 and Btk or Gap1^m, we studied the modulation

of Btk and Ras activity by endogenous G α 12-coupled receptors. Thromboxane A₂ receptors have been shown previously to be able to activate G α 12 in cells as well as in a reconstitution system containing purified receptors and G α 12 (refs 17, 18). We used MEG-01 human leukaemia cells as they express abundant thromboxane A₂ receptors¹⁹. Treatment of MEG-01 cells with the thromboxane A₂-receptor agonist U46619 induced the association of endogenous Btk and G α 12 (Fig. 7a). At the same time, U46619 stimulated Btk's kinase activity (Fig. 7b, c). As the thromboxane A₂ receptors are also able to couple to other G proteins¹⁸, we reasoned that, if G α 12 mediates stimulation of Btk by the thromboxane A₂ receptor, then transfection of the wild-type G α 12 should potentiate the U46619 stimulation. Transfection of wild-type G α 12 enhanced the U46619 stimulation of Btk activity, without affecting the Btk activity in the absence of U46619 (Fig. 7b, c). This is consistent with the idea that the thromboxane A₂ receptors, at least partly, generate active G α 12, which binds to and stimulates Btk.

Similarly, we studied the effect of thromboxane A₂ receptors on Gap1^m activity. Although Gap1^m might couple G α 12 signals to other pathways²⁰, we can at present study the effects of Gap1^m on Ras only. We determined whether endogenous thromboxane A₂ receptors could decrease the level of Ras-GTP stimulated by EGF and whether wild-type G α 12 could potentiate this suppression. U46619 treatment induced the co-immunoprecipitation of endogenous Gap1^m and G α 12 (Fig. 7d). At the same time, this U46619 treatment reduced the level of Ras-GTP induced by EGF (Fig. 7e). Transfection of wild-type G α 12 enhanced this suppression (Fig. 7e). Addition of U46619 had no effect on tyrosine phosphorylation of Shc induced by EGF (data not shown). These data indicate that G α 12, at least partly, mediates the inhibition of Ras by the thromboxane A₂ receptor, probably by interacting with Gap1^m.

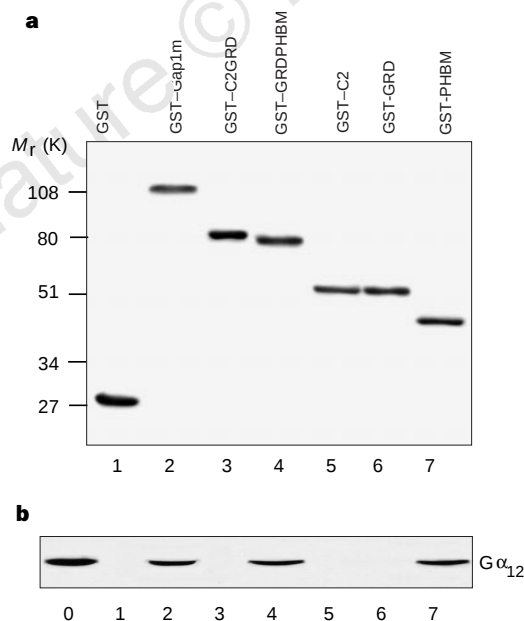


Figure 5 G α 12 association with Gap1^m. **a**, Coomassie staining of SDS-PAGE of purified GST-fusion proteins. **b**, Binding of G α 12 to the PH/BM region of Gap1^m. Lane 0 was loaded with 40 μ l whole-cell extract as control. Lanes 1–7 were loaded with GST-fusion proteins as in **a**.

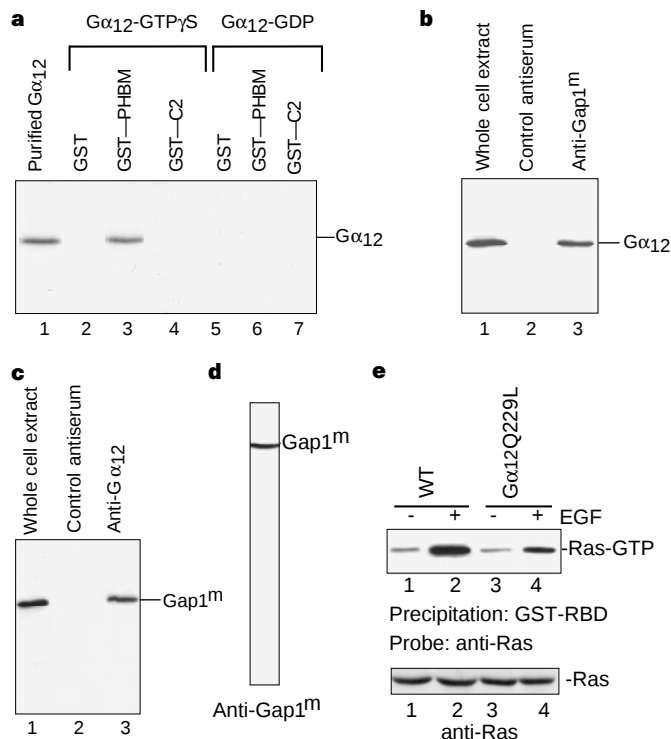


Figure 6 Effects of G α 12 on Ras signalling. **a**, direct interaction of purified G α 12-GTP γ S with a purified GST-fusion protein containing the PH/BM domain of Gap1^m *in vitro*. **b**, **c**, Co-immunoprecipitation of G α 12 with Gap1^m. **d**, Expression of Gap1^m in COS-7 cells as shown by Western blot. **e**, Top, overexpression of G α 12(Q229L) reduced the amount of Ras-GTP stimulated by EGF in COS-7 cells. In G α 12(Q229L)-transfected cells, the stimulation by EGF was reduced. Bottom, Western blot analysis of Ras protein in untransfected and transfected cells.

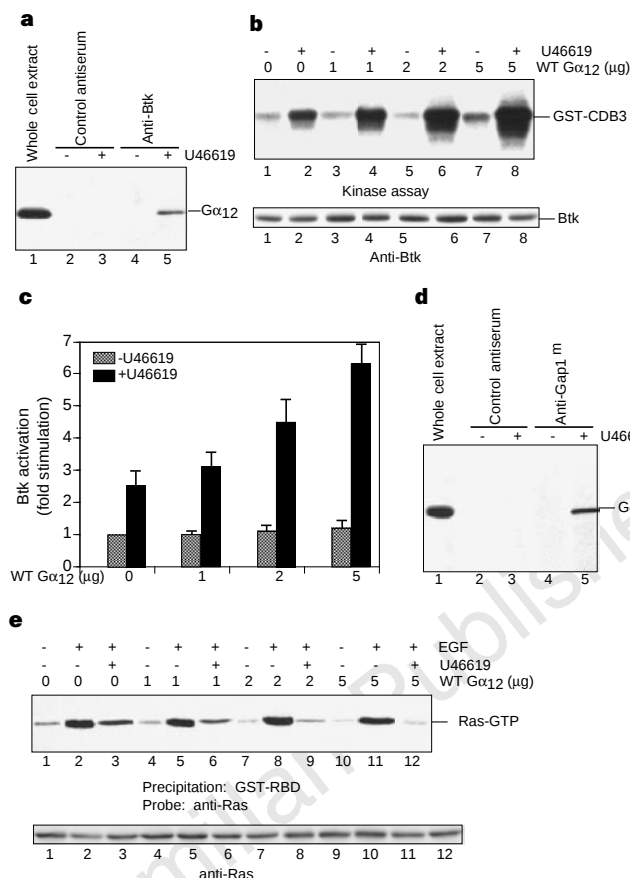


Figure 7 Regulation of the activity of Btk and Ras by endogenous thromboxane A_2 receptors. **a**, Stimulation of MEG-01 cells with a thromboxane- A_2 -receptor agonist, U46619, induced the association of endogenous Btk and $G\alpha_{12}$. **b**, Top, treatment with U46619 increased Btk activity towards GST-CDB3. Overexpression of the wild-type (WT) $G\alpha_{12}$ potentiated this stimulation. Bottom, western blot of Btk in transfected and untransfected MEG-01 cells. **c**, Quantification of stimulation of Btk kinase activity. Data represent mean $\pm$ s.d. of three

experiments. **d**, U46619 treatment of MEG-01 cells induced the co-immunoprecipitation of Gap1^m and $G\alpha_{12}$. **e**, Top, EGF activated Ras in MEG-01 cells, as indicated by the increase in the level of active Ras-GTP. This stimulation was decreased by pretreatment of cells with U46619. Overexpression of wild-type $G\alpha_{12}$ could potentiate this inhibitory effect of U46619. Bottom, western blot analysis of Ras proteins in transfected and untransfected cells.

We have shown that the α -subunit of the G_{12} protein can directly stimulate the activity of a tyrosine kinase, Btk, and a rasGAP, Gap1^m. This identifies the first specific direct effectors for $G\alpha_{12}$ proteins, and provides a route by which extracellular stimuli can modulate rasGAPs. Our results also show that there is a new type of modular interaction, between PH/BM domain and $G\alpha_{12}$, and reveal a direct link between heterotrimeric and monomeric G proteins through rasGAPs.

Methods

Protein purification. G-protein subunits were purified from recombinant baculovirus-infected Sf9 cells as described²¹. Recombinant Btk, Gap1^m and Ha-Ras were purified from *E. coli*⁷. Immunoprecipitation and western blot were done as described^{22,23}.

GST-fusion proteins. To construct GST-fusion proteins, DNA sequences corresponding to the indicated regions of human Btk complementary DNA or rat brain Gap1^m cDNA were amplified by the polymerase chain reaction (PCR) and subcloned into vector pGEX-2T (Pharmacia). Each construct was confirmed by DNA sequencing. GST-fusion proteins were expressed in BL21 cells and purified on glutathione-agarose beads.

In vitro binding assays. HEK-293 cells were transfected with pCDNA3- $G\alpha_{12}$ (Q229L) cDNA. Thirty-six hours later, cells were collected and whole-cell extracts prepared²⁴. 5 μ g purified GST-fusion proteins were incubated with $\sim$ 1 mg whole-cell extract for 2 h at 4°C, then washed three times with 1 $\times$ PBS + 0.1% Triton X-100 to remove unbound proteins. Bound $G\alpha_{12}$

was resolved on 10% SDS-PAGE, electroblotted onto nitrocellulose membrane, and identified with anti- $G\alpha_{12}$ antibody.

Affinity chromatographic binding assay. The *in vitro* affinity chromatographic binding assay was done as described²⁵. Briefly, 1 μ g purified GST-fusion protein and 0.05 μ g $G\alpha_{12}$ -GTP γ S or $G\alpha_{12}$ -GDP were incubated at room temperature for 5 min in 500 μ l total volume of buffer 0 (20 mM Tris-HCl, pH 8.0, 0.5 mM EDTA, 1 mM MgSO₄, 20% glycerol, 1 mM dithiothreitol (DTT), 0.2% NP-40 and 0.5 mM phenylmethylsulphonylfluoride (PMSF)). This was followed by incubation with 10 μ l glutathione-agarose beads for another 15 min. The reaction mixture was then loaded onto a glass-beads mini-column and washed three times with 1 ml wash buffer (buffer 0 with 100 mM KCl). Bound $G\alpha_{12}$ was eluted with 100 μ l elution buffer (buffer 0 with 1 M KCl and 1% deoxycholate), resolved by SDS-PAGE, and analysed by western blot.

Co-immunoprecipitation assay. Immunoprecipitations were done either with anti-Gap1^m monoclonal antibody (Transduction) or with anti-Btk polyclonal antibody (Santa Cruz Biotechnology). To avoid interference from immunoglobulin proteins on western blot analysis, co-immunoprecipitated proteins were eluted with buffer 0 with 500 mM KCl on ice for 30 min. Transient transfections were carried out with SuperFect Transfection Reagent (Qiagen). For detection of the interaction of endogenous proteins, MEG-01 human leukaemia cells (from ATCC) were treated with the thromboxane- A_2 -receptor agonist U46619 (Sigma) for 5 min before cells were collected.

Btk kinase assays. The Btk *in vitro* kinase assay was done as described⁷.

Immunocomplex kinase assay. The Btk immunocomplex kinase assay was done as described²³.

GTPase assay. GTPase assays were performed as described¹³. Briefly, Ha-Ras (7 μ l 1 μ M Ha-Ras in 20 mM HEPES, pH 7.3, 2 mM DTT, 1 mg ml⁻¹ BSA) was incubated for 5 min at 25 °C with [γ -³²P]GTP (6,000 Ci mmol⁻¹, NEN; 7 μ l 0.1 μ M diluted in 20 mM HEPES, pH 7.3, 2 mM DTT, 1 mM EDTA). [γ -³²P]GTP loading was stopped by the addition of 126 μ l 20 mM HEPES, pH 7.3, 2 mM MgCl₂, 2 mM DTT. For each GTPase assay, 5 μ l [γ -³²P]GTP-bound Ha-Ras was incubated for indicated times in the absence or presence of equal volumes of 1 nM Gap1^m in 20 mM HEPES, pH 7.3, 2 mM MgCl₂, 2 mM DTT, 100 μ M GTP, 1 mg ml⁻¹ BSA at 25 °C. In some reactions, certain concentrations of G proteins were included in the reactions. Measurement of free ³²P_i released from GTP hydrolysis was performed as described²⁶. Briefly, unreacted GTP was precipitated by the addition of 260 μ l acid-washed Norit-charcoal (5% in 20 mM phosphoric acid). After centrifugation (10 min), 100 μ l the supernatant was analysed for free ³²P_i by scintillation counting.

Detection of activated Ras. The assay for the interaction of activated Ras with the GST-fusion protein containing the Ras-binding domain (RBD) of Raf-1 was performed as described¹⁶. COS-7 cells or MEG-01 cells, untransfected or transfected with G α 12 plasmid DNA, were grown for 48 h and then starved for 18 h. 3 μ M U46619 was added for 10 min before the addition of 50 nM EGF for 5 min. Whole-cell extracts were prepared. Activated Ras-GTP was precipitated with 20 μ l glutathione-agarose beads with GST-RBD fusion protein (~10 μ g). After 12% SDS-PAGE, Ras was detected with anti-Ras antibodies.

Received 22 June; accepted 13 August 1998.

- Gilman, A. G proteins: transducers of receptor-generated signals. *Annu. Rev. Biochem.* **56**, 615–649 (1987).
- Bourne, H. R., Sanders, D. A. & McCormick, F. The GTPase superfamily: a conserved switch for diverse cell functions. *Nature* **348**, 125–132 (1990).
- Simon, M. I., Strathmann, M. P. & Gautum, N. Diversity of G proteins in signal transduction. *Science* **252**, 802–808 (1991).
- Clapham, D. E. & Neer, E. J. G protein $\beta\gamma$ subunits. *Annu. Rev. Pharmacol. Toxicol.* **37**, 167–203 (1997).
- Strathmann, M. P. & Simon, M. I. G α 12 and G α 13 subunits define a fourth class of G protein α subunits. *Proc. Natl Acad. Sci. USA* **88**, 5582–5586 (1991).
- Dhanasekaran, N. & Dermott, J. M. Signaling by the G12 class of G proteins. *Cell. Signal.* **8**, 235–245 (1996).
- Bence, K., Ma, W., Kozasa, T. & Huang, X.-Y. Direct stimulation of Bruton's tyrosine kinase by Gq-protein α -subunit. *Nature* **389**, 296–299 (1997).
- Vihinen, M. et al. BTKbase: a database of XLA-causing mutations. *Immunol. Today* **16**, 460–465 (1995).
- Vihinen, M., Nilsson, L. & Smith, C. I. E. Tec homology (TH) adjacent to the PH domain. *FEBS Lett.* **350**, 263–265 (1994).
- Maekawa, M., Nakamura, S. & Hattori, S. Purification of a novel ras GTPase-activating protein from rat brain. *J. Biol. Chem.* **268**, 22948–22952 (1993).
- Maekawa, M. et al. A novel mammalian Ras GTPase-activating protein which has phospholipid-binding and Btk homology regions. *Mol. Cell. Biol.* **14**, 6879–6885 (1994).
- Gaul, U., Mardon, G. & Rubin, G. M. A putative Ras GTPase activating protein acts as a negative regulator of signaling by the Sevenless receptor tyrosine kinase. *Cell* **68**, 1007–1019 (1992).
- Cullen, P. J. et al. Identification of a specific Ins(1,3,4,5)P₄-binding protein as a member of the GAP1 family. *Nature* **376**, 527–530 (1995).
- Yamamoto, T., Matsui, T., Nakafuku, M., Iwamatsu, A. & Kaibuchi, K. A novel GTPase-activating protein for R-Ras. *J. Biol. Chem.* **270**, 30557–30561 (1995).
- Voyno-Yasenetskaya, T. A., Faure, M. P., Ahn, N. G. & Bourne, H. R. G α 12 and G α 13 regulate extracellular signal-regulated kinase and c-Jun kinase pathways by different mechanisms in COS-7 cells. *J. Biol. Chem.* **271**, 21081–21087 (1996).
- Taylor, S. J. & Shalloway, D. Cell cycle-dependent activation of Ras. *Curr. Biol.* **6**, 1621–1627 (1996).
- Offermanns, S., Laugwitz, K.-L., Spicher, K. & Schultz, G. G proteins of the G12 family are activated via thromboxane A₂ and thrombin receptors in human platelets. *Proc. Natl Acad. Sci. USA* **91**, 504–508 (1994).
- Harhammer, R. et al. Distinct biochemical properties of the native members of the G12 G-protein subfamily. *Biochem. J.* **319**, 165–171 (1996).
- Hirata, M. et al. Cloning and expression of cDNA for a human thromboxane A₂ receptor. *Nature* **349**, 617–620 (1991).
- McCormick, F. Activators and effectors of ras p21 proteins. *Curr. Opin. Genet. Dev.* **4**, 71–76 (1994).
- Kozasa, T. & Gilman, A. G. Purification of recombinant G proteins from Sf9 cells by hexahistidine tagging of associated subunits. *J. Biol. Chem.* **270**, 1734–1741 (1995).
- Wan, Y., Kurosaki, T. & Huang, X.-Y. Tyrosine kinases in activation of the MAP kinase cascade by G-protein-coupled receptors. *Nature* **380**, 541–544 (1996).
- Wan, Y. et al. Genetic evidence for a tyrosine kinase cascade preceding the mitogen-activated protein kinase cascade in vertebrate G protein signalling. *J. Biol. Chem.* **272**, 17209–17215 (1997).
- Langhans-Rajasekaran, S. A., Wan, Y. & Huang, X.-Y. Activation of Tsk and Btk tyrosine kinases by G protein $\beta\gamma$ subunits. *Proc. Natl Acad. Sci. USA* **92**, 8601–8605 (1995).
- Zhang, J. J. et al. Two contact regions between Stat1 and CBP/p300 in interferon γ signaling. *Proc. Natl Acad. Sci. USA* **93**, 15092–15096 (1996).
- Schwer, B. & Guthrie, C. PRP16 is an RNA-dependent ATPase that interacts transiently with the spliceosome. *Nature* **349**, 494–499 (1991).

Acknowledgements. We thank S. Taylor and D. Shalloway for the pGEX-RBD plasmid; M. Liu and M. Simon for the G α 12 plasmid; and M. Gershengorn, L. Levin, W. Lowry and T. Maack for reading the manuscript. Some of the Sf9 cells were cultured in the National Cell Culture Center. This work was supported by grants from the NIH, the NSF, and the American Heart Association. X.-Y.H. is a Beatrice E. Parvin Investigator of the American Heart Association New York City affiliate.

Correspondence and requests for materials should be addressed to X.-Y.H. (e-mail: xyhuang@mail.med.cornell.edu).

Direction determination in the minus-end-directed kinesin motor ncd

Elena P. Sablin*, Ryan B. Case*, Shirleko C. Dai*, Cynthia L. Hart*, Aaron Ruby*, Ronald D. Vale*† & Robert J. Fletterick*

* Departments of Biochemistry/Biophysics and Pharmacology and

† Howard Hughes Medical Institute, University of California, San Francisco, California 94143, USA

Motor proteins of the kinesin superfamily transport intracellular cargo along microtubules. Although different kinesin proteins share 30–50% amino-acid identity in their motor catalytic cores, some move to the plus end of microtubules whereas others travel in the opposite direction^{1,2}. Crystal structures of the catalytic cores of conventional kinesin (a plus-end-directed motor involved in organelle transport) and ncd (a minus-end-directed motor involved in chromosome segregation) are nearly identical^{3,4}; therefore, the structural basis for their opposite directions of movement is unknown. Here we show that the ncd 'neck', made up of 13 class-specific residues next to the superfamily-conserved catalytic core, is essential for minus-end-directed motility, as mutagenesis of these neck residues reverses the direction of ncd motion. By solving the 2.5 Å structure of a functional ncd dimer, we show that the ncd neck (a coiled-coil) differs from the corresponding region in the kinesin neck (an interrupted β -strand)^{5,6}, although both necks interact with similar elements in the catalytic cores. The distinct neck architectures also confer different symmetries to the ncd and kinesin dimers and position these motors with appropriate directional bias on the microtubule.

We determined the structure of an ncd homodimer (E281–K700, refs 7, 8; Fig. 1a, b) in a complex with Mg-ADP by using molecular-replacement methods (Table 1). The model of the ncd homodimer includes 740 amino-acid residues (amino acids (aa) 303–672 from each chain). The motor heads are connected through their 'necks', 13 residues that precede the first β -strand of the ncd catalytic core and are defined by high conservation among only the carboxy-terminal motor proteins of the kinesin superfamily (those kinesin proteins with motor domains at their C-terminal ends)² (Fig. 1a). The neck is entirely helical and partners symmetrically in the dimer to form a parallel coiled-coil (Fig. 1b). The conserved neck connects to the less conserved coiled-coil 'stalk' domain without any intervening loops. The coiled-coil helices of the neck are connected to the motor heads in each monomer through an abrupt turn that contains the conserved residue G347 at its apex. The structure of

Table 1 Data collection and model building and refinement

Space group: P6 ₁ 22	
a = 123.0 Å, b = 123.0 Å, c = 121.1 Å	
Data collection	
Resolution (Å)	2.5
R _{sym} (%)	6.8
Unique reflections	18,795
Observed reflections	190,358
Completeness (%)	98.6
Refinement (20.0–2.5 Å)	
R _{cyst} (%)	22.6
R _{free} (%)	27.0
r.m.s. bond length (Å)	0.007
r.m.s. bond angle (degrees)	1.44
Average B-factor (Å ²)	23.5
r.m.s., root mean square.	

each catalytic core is unchanged from the previously determined structure of the ncd monomer catalytic core⁴. An exception is loop L11 (Fig. 1b), which was possibly ordered by crystal packing in the monomer, but is disordered in the dimer structure.

The ncd dimer structure has two-fold symmetry, which is set by the coiled-coil domain. The two-fold symmetry of the ncd dimer seen in the crystal structure also exists in solution, as confirmed by low-angle X-ray scattering of the ncd-ADP complex (R. Mendelson, manuscript in preparation). However, the symmetrical structure of the ncd dimer must represent only one state of the motor during its mechanochemical cycle—the microtubule-dissociated ADP-bound state. Electron microscopic images of the ncd dimer

bound to microtubules show that the two motor heads, even though they are still closely apposed as in the crystal structure, are not related by two-fold symmetry^{9–11}. Therefore, a symmetry-breaking conformational change in the ncd dimer must be associated with binding of the motor to the microtubule and the subsequent release of ADP from the microtubule-bound head. The rotation of one head in the dimer crystal structure by ~25° about the two-fold axis could explain the position of the detached head in the cryo-electron microscopic images.

The neck and stalk coiled-coil domains interact with the catalytic core through an extensive network of interactions which creates a continuous interface of ~690 Å². The interacting core residues cluster mainly in the C-terminal part of helix α1 and loops L6, L10 and L13 (Fig. 2a, b). Six of the residues on ncd's catalytic core that form the interfacial surface are conserved only among the C-terminal, minus-end-directed kinesin motors (Q420, S421, D424, G472, W473 and K640)⁴. However, the residue Y426 is conserved throughout the kinesin superfamily⁴ (it is equivalent to Y77 in human conventional kinesin). Six amino acids on the complementary interface of the ncd neck region (R335, K336, H339, N340, D344 and R346) are highly conserved among the C-terminal kinesin motors².

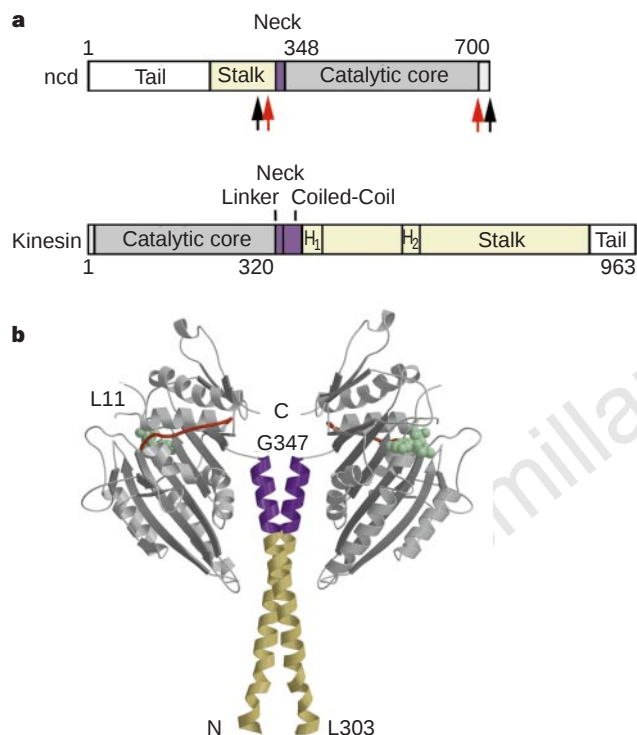


Figure 1 Structure of the ncd dimer. **a**, Domain organization of ncd and kinesin homodimers. The polypeptide chain of the monomer is organized into four domains (right to left, C to N terminus, in ncd; left to right in kinesin). First, there is a globular catalytic core (grey), which contains the ATP- and microtubule-binding sites and is conserved throughout the kinesin superfamily. The ncd catalytic core is about 40% identical in sequence to kinesin's catalytic core but, in contrast to in kinesin, it is located at the C terminus of the ncd polypeptide chain. Second, there is a 'neck' region (purple), which is adjacent to the catalytic core and is defined by class-specific sequence conservation¹². The neck of ncd contains 13 class-conserved amino acids (R335–G347) that precede the first β-strand of the catalytic core. Kinesin's conserved neck (~35 residues) emerges from the C terminus of the catalytic core and consists of two distinct regions. The first ten residues, termed the 'neck linker', are highly conserved among all plus-end-directed motors and interact with the catalytic core. The subsequent region forms a coiled-coil ('neck coiled-coil'), does not interact with the core, and is conserved selectively among the conventional kinesin subfamily¹². Third, there is an α-helical 'stalk' domain (yellow) that enhances dimer formation through an extended coiled-coil. H₁ and H₂ indicate flexible hinge regions. Finally, there is a small C-terminal 'tail' or attachment domain (white). The black and red arrows delineate the characterized ncd fragment and the visible part of the structure, respectively. **b**, View perpendicular to the dyad axis showing the course of the polypeptide chains of ncd; the dimer has dimensions of ~95 × 105 × 65 Å, with the two catalytic cores (heads) positioned 45 Å from each other. The colouring scheme is consistent with **a**. ADP is shown in the active site in green. One of the important secondary-structure elements involved in microtubule-binding (loop L12) is shown in orange.

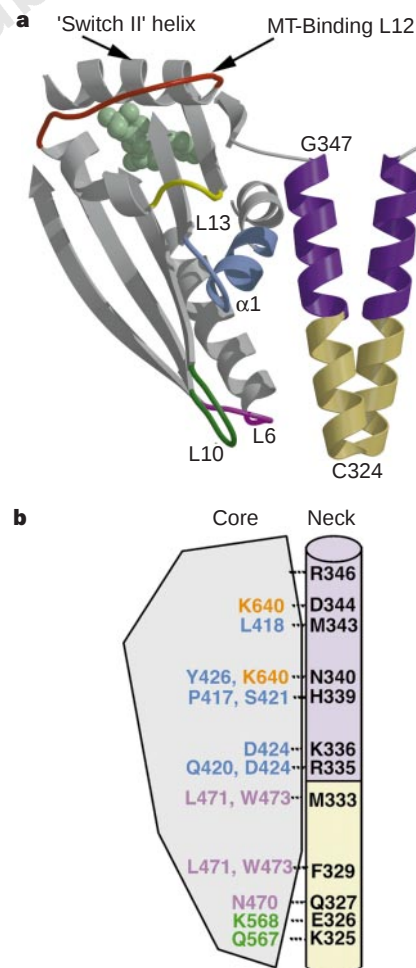


Figure 2 Architecture of the ncd neck-core interface. **a**, Interactions of the neck with the motor heads: the neck is shown in purple; regions of the core that contact the neck are shown in blue (α1), magenta (L6), green (L10) and yellow (L13); and the stalk domain is golden. Loop L13 is next to the predicted microtubule (MT)-binding site of ncd (loop L12 and helix α5). Bound ADP is shown as green spheres. **b**, The interdomain contacts; the colouring scheme for the interacting residues in the core is consistent with **a**.

Table 2 Microtubule gliding velocities of ncd mutants

Construct	Motor direction	Velocity (% wild-type)
Wild-type	Minus	100 ± 5
E567A/K568A	Minus	99 ± 11
H339A/N340A	Minus	76 ± 11
G636A/G637A	Minus	79 ± 7
Y426A	Minus	57 ± 7
Q420A/S421A	Minus	44 ± 14
R335A/K336A	Minus	49 ± 9
K640A	Minus	34 ± 0.2
Q420A/S421A/Y426A	Minus	10 ± 1
NCD-ran12 (aa335 RKE LHNTVMDLR aa346 to ESGAKQGEKGES)	Plus	2 ± 0.3

The gliding velocity of wild-type ncd-GFP was $8.4 \pm 0.4 \mu\text{m min}^{-1}$. Values are the mean $\pm$ s.d. of the average velocities obtained from two to three independent protein preparations (50 microtubules per preparation). The affinity of ncd-ran12 for microtubules in the motility assay was lower than that of wild-type, and fewer microtubules were observed on the ncd-ran12 surface.

To study the function of conserved residues in the ncd neck–core interface, we prepared nine mutants and tested their motility (Table 2). The most drastic mutation replaced 12 neck residues (aa 335–346) with random, hydrophilic residues (ESGAKQGEKGES; named ncd-ran12), which should completely disrupt the neck–core interaction and the coiled-coil architecture of the ncd neck. Most of the mutations had small effects on motility (less than twofold change). However, mutant K640A and the triple mutant Q420A/S421A/Y426A exhibited 3- and 10-fold, respectively, lower microtubule gliding velocities (minus-end-directed) than wild-type ncd. Surprisingly, ncd-ran12 moved slowly (at $0.16 \mu\text{m min}^{-1}$) towards the microtubule plus end, that is, in the opposite direction to the wild-type ncd motor. As ncd-ran12 lacks its conserved neck, its motility may be generated through a small plus-end-directed conformational change associated with the catalytic core, as may also occur for a similar neck-replaced kinesin mutant (R. Case *et al.*, manuscript in preparation).

Although the Q420A/S421A/Y426A and ncd-ran12 mutants generated slow motility, their ATPase turnover was nearly normal, indicating that nucleotide hydrolysis had become uncoupled from motility (k_{cat} values of 1.5 ± 0.1 , 0.9 ± 0.2 and 1.4 ± 0.2 ATP molecules hydrolysed per head per second by wild-type ncd, Q420A/S421A/Y426A and ncd-ran12 respectively; mean $\pm$ s.d. of values from 2–3 protein preparations). These mutants also exhibited lower affinity for microtubules, as reflected by their elevated K_m (microtubules) for ATPase activation (1.0 ± 0.1 , 1.7 ± 0.7 , $19.4 \pm 7.4 \mu\text{M}$ tubulin dimer for wild-type ncd, Q420A/S421A/Y426A and ncd-ran12, respectively). These results indicate that ncd motor activity can be uncoupled from ATP turnover by mutating conserved residues at the neck–core interface and that an intact neck is essential for generating minus-end-directed motion.

There are major differences between the dimeric ncd and conventional kinesin⁵ crystal structures that may be relevant to understanding their opposite directions of motion (Fig. 3). First, the arrangements of the heads differ in the two motor dimers. The kinesin heads do not show two-fold symmetry but are related by a $\sim 120^\circ$ rotation and also are positioned further apart from each other than the heads of ncd. The different symmetries of the kinesin and ncd dimers appear to be determined by their distinct neck architectures. In contrast to the continuous coiled-coil of the ncd neck, the kinesin neck consists of two short β -strands ($\beta 9$ and $\beta 10$; termed the neck linker) followed by the coiled-coil helix $\alpha 7$ (neck coiled-coil), all of which are connected by short loops⁵ (Figs 1 and 3). According to the crystal structures, if the bound ncd and kinesin heads are positioned similarly on a microtubule, as indicated by cryo-electron microscopic experiments¹², then the unbound kinesin head would point towards the microtubule plus end, whereas the unbound ncd head would be tilted towards the minus end (Fig. 3). This analysis indicates that the different symmetries of the ncd and

kinesin heads in dimeric motors position the detached heads in the direction of motor movement. The different positioning of the two heads by their necks, which is consistent with cryo-electron microscopic studies of microtubule-bound ncd and kinesin dimers^{9,10}, should contribute to the opposite direction of movement of these two motors.

Despite their different sequences and structures, the ncd and kinesin necks both follow the same path and interact with the same secondary-structure elements in the catalytic core (mainly the $\alpha 1/\beta 3$ turn and loop L13). This similarity was not expected, because the helical neck of ncd emerges before the amino terminus of the catalytic core, whereas the kinesin neck follows the C terminus of the core. In both cases, the neck–core contacts are near the main microtubule-binding loop, L12 (refs 11, 13), and the ‘switch II’ helix $\alpha 4$ (Fig. 2a), which are thought to change conformation during the nucleotide cycle⁴. Two major predictions can be made on the basis of this observation. First, the catalytic core could convey information about its nucleotide- and microtubule-binding state in a similar manner (through $\alpha 1$ and L13) to the necks of both motors. However, the necks, because of their different architectures, must respond and transmit this information to the unattached, partner heads differently, so that they can move in opposite directions along the microtubule. Second, the similar core–neck communication pathway in both motors could explain the co-functioning of the kinesin neck with the ncd catalytic core in chimaeric proteins^{14,15}.

Our structural and motility data indicate that the helical neck domain of ncd, with its sophisticated interactions with the core, is necessary for generating minus-end-directed motion. How does the ncd neck function in the force-generating cycle? Although the

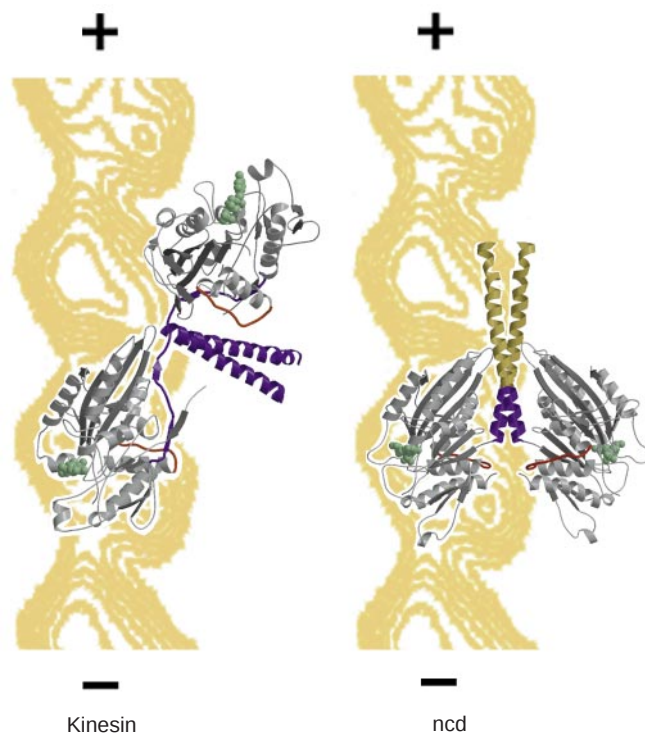


Figure 3 Model showing the ncd and kinesin dimer structures docked onto a tubulin protofilament. The colouring scheme for the parts of the kinesin and ncd dimers is consistent with Fig. 1. The bound ncd and kinesin heads are positioned similarly, with loop L12 (red) docked onto the tubulin (background). Because of the distinct architectures of the kinesin and ncd necks, the unbound kinesin head points towards the plus end, whereas the unbound ncd head is tilted towards the minus end of the protofilament.

helical structure of the neck may be stabilized by its specific interactions with the catalytic core in the ncd-dimer-ADP complex, the neck may be able to adopt other conformations during the enzymatic cycle. In support of this idea, predictive analysis indicates that the ncd neck residues (R335–G347) are not expected to form either a stable α -helix (PHD program)¹⁶ or a coiled-coil (COILS program)¹⁷, and a synthetic peptide corresponding to the ncd neck (L328–R348) is not helical in solution (B. Tripet and R. S. Hodges; and T. Shimizu, M. Itoh, H. Morii and M. Tanokura, personal communications). Moreover, the neck is disordered and not bound to the core in the ncd monomer structure⁴. The polar character of the residues at the neck–core interface also indicates that this region is designed to be dynamic. We propose that the ncd force-generating mechanism includes a conformational change of the neck, possibly involving a partial or complete loss of its α -helical structure. Conformational transitions of the neck could be initiated from the nearby nucleotide-binding and microtubule-binding sites (through the switch II and L12 regions, respectively) and transmitted through the neck–core interface. The net result of this conformational transition would be expected to produce a force vector directed towards the microtubule minus end and, possibly, a torque, which has been seen in ncd motility assays¹⁸. In contrast, the kinesin neck appears to amplify motion towards the microtubule plus end (R. Case *et al.*, manuscript in preparation). In the ncd and kinesin dimers, the oppositely directed actions of the neck in the microtubule-bound head would then be further amplified by the detached partner heads, which are positioned with appropriate directional bias in the two motors. □

Methods

Expression, purification and crystallization of the ncd dimer. A DNA fragment encoding residues E281–K700 of the ncd motor protein^{7,8} was synthesized by the polymerase chain reaction (PCR) and cloned into the translation vector pHB40P. The recombinant plasmid was transformed into BL21(DE3)pLysE host cells (Novagen) for expression. Cells were grown at 24 °C in LB media supplemented with ampicillin (0.1 mg ml⁻¹) and induced by addition of 0.2 mM of isopropyl- β -D-thiogalactoside (IPTG) for 8 h. Packed, induced cells were resuspended in PB buffer (10 mM PIPES, pH 7.2, 100 mM NaCl, 2 mM MgCl₂, 1 mM EGTA, 1 mM dithiothreitol (DTT), and protease inhibitors) with 0.1 mg ml⁻¹ DNase I, and lysed using a French pressure cell press (American Instrument). The supernatant from a 60,000g centrifugation for 1 h was loaded onto a SP–Sephacrose FF (Pharmacia) column equilibrated in PB. A linear NaCl gradient (0.1–0.3 M) was used to elute the ncd, and the pooled fractions were dialysed against TB buffer (10 mM Tris-HCl, pH 7.4, 80 mM NaCl, 2 mM MgCl₂, 1 mM DTT and protease inhibitors) and loaded onto a MonoQ HR10/10 column (Pharmacia) equilibrated in TB. A linear NaCl gradient (0.1–0.3 M) was applied to elute ncd, which was then concentrated to ~20 mg ml⁻¹ using a Centricon-30 (Amicon) and used fresh for crystallization. The purified protein showed a microtubule-activated ATPase activity of 0.5 ATP molecules per head per second. Crystals were obtained by vapour diffusion using 20 μ l sitting drops. The mother liquor contained protein at about 20 mg ml⁻¹, 2 mM ADP, 10 mM MgCl₂, 100 mM NaCl, 700 mM Li₂SO₄, 1 mM EGTA and 1 mM DTT in 20 mM HEPES, pH 7.5. The reservoir was 1.4 M Li₂SO₄, 10 mM MgCl₂, 1 mM EGTA and 1 mM DTT in 20 mM HEPES, pH 7.5. Crystals appeared after 2 weeks at +4 °C, and were of the hexagonal space group P6₁22, with one ncd monomer plus bound Mg-ADP per asymmetric unit.

Data collecting, model building and refinement. X-ray diffraction data to 2.5 Å resolution were measured at –170 °C using 30% (v/v) glycerol in precipitant solution as a cryosolvent. The data were collected at Stanford Synchrotron Radiation Laboratory beamline 7-1 ($\lambda = 1.08$ Å), integrated using DENZO and scaled with SCALEPACK (Z. Otwinowski and W. Minor). The structure of the ncd dimer was determined to 2.5 Å by molecular replacement (AMoRe) using atomic coordinates for the ncd monomer (residues R335–K700)⁴. An electron-density map based on coefficients $2F_o - F_c$ was calculated from the phases of the initial model. The additional parts of the structure were built and placed in visible density using the program O (T. A. Jones and M. Kjeldgaard). Later stages of model building included conjugate

gradient minimization and refinement by simulated annealing using X-PLOR¹⁹. The entire structure was checked and rebuilt using annealed omit maps. The current atomic model includes 740 amino-acid residues together with 2 Mg-ADP complexes, 54 water molecules and 4 sulphate anions. The final maps based on the refined coordinates, contain electron density for amino-acid residues from L303 to M672. Electron density is missing for the N-terminal amino-acid residues 281–302, loop L11 (aa 588–596) and C-terminal residues 673–700.

Mutagenesis. Site-directed mutagenesis using Stratagene's QuikChange protocol was done in a previously characterized ncd–green fluorescent protein (GFP) construct which consisted of a His₆ tag at the N terminus of GFP, a 2-amino-acid Gly–Thr linker (*Kpn*I site) at the C terminus of GFP, followed by ncd residues 236–700 (refs 7, 15). The sequences of all constructs were verified. Protein was expressed in bacteria, applied to Ni-NTA resin (Qiagen), and further purified on a mono-S or high trap-SP column (Pharmacia) as described¹⁵. Protein purity was in the range 50–85%, with most contaminants representing degradation products. Protein concentration was measured as described¹³. Proteins were frozen (10% sucrose added) and stored in liquid nitrogen.

Microtubule-stimulated ncd ATPase was measured in 15 mM NaMOPS (pH 7), 3 mM NaCl, 2 mM MgCl₂, 1 mM EGTA, 1 mM DTT, 1 mM ATP and 10 μ M paclitaxel using a malachite green assay²⁰. Gliding assays for ncd–GFP were performed using rhodamine-labelled microtubules and fluorescence microscopy as described¹⁵. For slow-moving proteins (<1 μ m min⁻¹), we used polarity-marked, rhodamine-labelled microtubules²¹ to ensure that all motion was unidirectional and that stage drift did not account for the motion; the illumination was shuttered to reduce damage from free radicals. We quantified 50 microtubule gliding speeds and directionalities for each protein preparation.

Received 14 May; accepted 2 September 1998.

- Hirokawa, N. Kinesin and dynein superfamily proteins and the mechanism of organelle transport. *Science* **279**, 519–526 (1998).
- Vale, R. D. & Fletterick, R. J. The design plan of kinesin motors. *Annu. Rev. Cell Dev. Biol.* **13**, 745–777 (1997).
- Kull, F. J., Sablin, E. P., Lau, R., Fletterick, R. J. & Vale, R. D. Crystal structure of the kinesin motor domain reveals a structural similarity to myosin. *Nature* **380**, 555–559 (1996).
- Sablin, E. P., Kull, F. J., Cooke, R., Vale, R. D. & Fletterick, R. J. Crystal structure of the motor domain of the kinesin-related motor ncd. *Nature* **380**, 555–559 (1996).
- Kozelski, F. *et al.* The crystal structure of dimeric kinesin and implications for microtubule-dependent motility. *Cell* **91**, 985–994 (1997).
- Sack, S. *et al.* X-ray structure of motor and neck domains from rat brain kinesin. *Biochemistry* **36**, 16155–16165 (1997).
- McDonald, H. B. & Goldstein, L. S. Identification and characterization of a gene encoding a kinesin-like protein in *Drosophila*. *Cell* **61**, 991–1000 (1990).
- Endow, S. A., Henikoff, S. & Soler, N. L. Mediation of meiotic and early mitotic chromosome segregation in *Drosophila* by a protein related to kinesin. *Nature* **345**, 81–83 (1990).
- Hirose, K., Lockhard, A., Cross, R. A. & Amos, L. A. Three-dimensional cryoelectron microscopy of dimeric kinesin and ncd motor domains on microtubules. *Proc. Natl. Acad. Sci. USA* **93**, 9539–9544 (1996).
- Arnal, I., Metoz, F., DeBonis, S. & Wade, R. H. Three-dimensional structure of functional motor proteins on microtubules. *Curr. Biol.* **6**, 1265–1270 (1996).
- Sosa, H. *et al.* A model for the microtubule–Ncd motor protein complex obtained by cryo-electron microscopy and image analysis. *Cell* **90**, 217–224 (1997).
- Amos, L. A. & Hirose, K. The structure of microtubule–motor complexes. *Curr. Opin. Cell Biol.* **9**, 4–11 (1997).
- Woehlke, G. *et al.* Microtubule interaction site of the kinesin motor. *Cell* **90**, 207–216 (1997).
- Henningsen, U. & Schliwa, M. Reversal in the direction of movement of a molecular motor. *Nature* **389**, 93–96 (1997).
- Case, R. B., Pierce, D. W., Hom-Booher, N., Hart, C. L. & Vale, R. D. The directional preference of kinesin motors is specified by an element outside of the motor catalytic domain. *Cell* **90**, 959–966 (1997).
- Rost, B. & Sander, C. PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol.* **266**, 525–539 (1996).
- Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
- Walker, R. A., Salmon, E. D. & Endow, S. A. The *Drosophila* claret segregation protein is a minus-end directed motor molecule. *Nature* **347**, 780–782 (1990).
- Brunger, A. T. *X-PLOR Version 3.1. A System for X-ray Crystallography and NMR* (Yale Univ. Press, New Haven, 1992).
- Kodama, T., Fukui, K. & Kometani, K. The initial phosphate burst in ATP hydrolysis by myosin and subfragment 1 as studied by a modified Malachite Green method for determination of organic phosphate. *J. Biochem.* **99**, 1465–1472 (1986).
- Hyman, A. A. Preparation of marked microtubules for the assay of the polarity of microtubule-based motors by fluorescence. *J. Cell Sci. Suppl.* **14**, 125–127 (1991).

Acknowledgements. We thank M. Matuska for technical assistance, and J. Somoza and C. Sindelar for comments on manuscript. E.P.S. was supported by a postdoctoral fellowship from the Leukemia Society of America. This work was supported in part by a program project grant from the NIH.

Correspondence and requests for materials should be addressed to R.D.V. (e-mail: vale@phy.ucsf.edu). X-ray crystallographic coordinates have been deposited in the Protein Data Bank under accession numbers 2ncd and r2ncdsf.

Unfolded conformations of α -lytic protease are more stable than its native state

Julie L. Sohl^{*†‡}, Sheila S. Jaswal^{‡§} & David A. Agard^{*§}

^{*} Graduate Group in Biophysics and the Howard Hughes Medical Institute,

[§] Department of Biochemistry and Biophysics, University of California at San Francisco, San Francisco, California 94143-0448, USA

[‡] These authors contributed equally to this work

[†] Present address: Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720, USA.

α -Lytic protease (α LP), an extracellular bacterial protease, is synthesized with a large amino-terminal pro-region that is essential for its folding *in vivo* and *in vitro*^{1,2}. In the absence of the pro-region, the protease folds to an inactive, partially folded state, designated 'I'. The pro-region catalyses protease folding by directly stabilizing the folding transition state (>26 kcal mol⁻¹) which separates the native state 'N' from I^{1,3}. Although a basic tenet of protein folding is that the native state of a protein is at the minimum free energy⁴, we show here that both the I and fully unfolded states of α LP are lower in free energy than the native

state. Native α LP is thus metastable: its apparent stability derives from a large barrier to unfolding. Consequently, the evolution of α LP has been distinct from most other proteins: it has not been constrained by the free-energy difference between the native and unfolded states, but instead by the size of its unfolding barrier.

The α LP N state is compact, with a well ordered hydrophobic core, and is stable to chemical and thermal denaturation (see ref. 5 and below). In contrast, the I state has some secondary structure but little or no tertiary structure¹. Gel filtration¹ and analytical ultracentrifugation (Fig. 1a) indicate that I is a greatly expanded monomer. Not surprisingly, I is temperature-sensitive¹ and it is stabilized by less than 1 kcal mol⁻¹ relative to the unfolded state 'U' (Fig. 1b). The conventional view of the forces that drive protein folding⁴ would place N at the minimum Gibbs free energy. In principle, however, it is possible for a protein to function even if its native state is not at the free-energy minimum, provided that the kinetic barrier is sufficient to prevent unfolding over the protein's functional lifetime. We now demonstrate that this is the case for α LP.

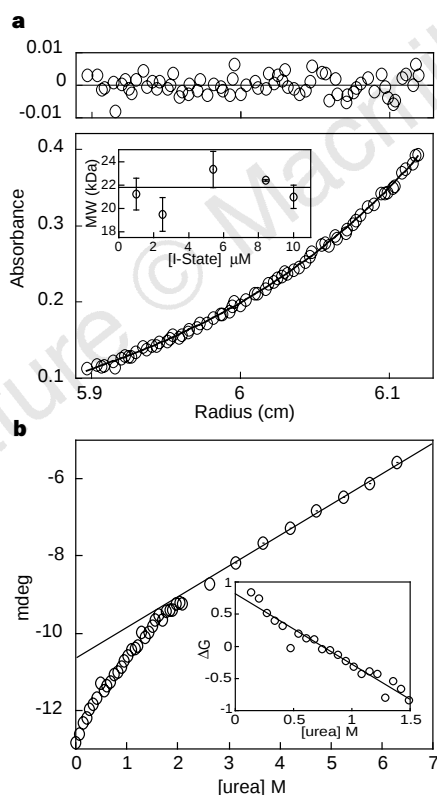


Figure 1 Characterization of the α LP I state. **a**, Sedimentation equilibrium analysis of 10 μ M α LP I state with monomer fit and residuals. As shown in the inset, the I state remains monomeric, with an average M_r of 21.8K over at least a 10-fold concentration range. **b**, Urea denaturation of the I state followed by its circular dichroism signal at 225 nm. Inset shows the free energies for the I $\rightarrow$ U transition as a function of urea, calculated using the displayed linear unfolded baseline and by assigning the I-state baseline to the 0M urea value. The calculated $\Delta G_{I \rightarrow U}$ at 0M urea must be considered a maximum value.

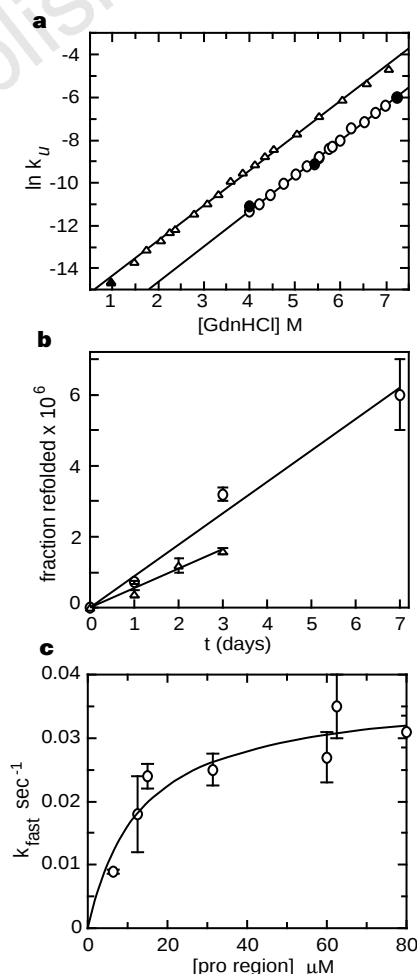


Figure 2 The rates of α LP unfolding and refolding. **a**, The logarithm of the unfolding rates at 4°C (circles) and 25°C (open triangles) as a function of denaturant (guanidinium (Gdn) hydrochloride) concentration. Open symbols denote the rates of tertiary structure loss of S195A observed using tryptophan fluorescence; the filled triangle represents the rate of activity loss of the wild-type protease; filled circles show the rates of secondary structure loss in S195A monitored by circular dichroism at 225 nm. **b**, Pro-region-independent folding of the I to N states as a function of time at 4°C (circles) and 25°C (triangles). **c**, The concentration dependence of the fast phase rates of pro-region-catalysed I to N folding.

The large folding barrier prevents equilibration of I and N on any practical time scale; thus the relative stability of these states is addressed indirectly through the rates of folding (k_f) and unfolding (k_u). At 25 °C, unfolding of the protease in 1 M guanidinium-HCl can be monitored by loss of activity. With increasing denaturant concentration, measurements of wild-type protease unfolding are complicated by autolysis, so we used a variant of the protease in which the active-site serine is replaced by alanine (Ser 195 → Ala, S195A, where the residue numbering for α LP is based on homology to chymotrypsin³) to measure the rates of unfolding as a function of denaturant at 4 °C and 25 °C. During unfolding, secondary and tertiary structure are lost simultaneously and the spectroscopic data are best fitted by a single exponential function of time. The observed rates are plotted as a function of guanidinium-HCl concentration in Fig. 2a. By linear extrapolation to zero concentration of guanidinium, the rate of unfolding is $1.8 \times 10^{-8} \pm 0.2 \times 10^{-8} \text{ s}^{-1}$ at 4 °C ($t_{1/2} = 1.2 \text{ y}$) and $1.2 \times 10^{-7} \pm 0.6 \times 10^{-7} \text{ s}^{-1}$ at 25 °C ($t_{1/2} = 70 \text{ d}$). Urea denaturation at 4 °C yields the same extrapolated rate within error (data not shown), supporting the use of a linear extrapolation to determine k_u .

To measure k_f , we incubated solutions of I at 4 °C and 25 °C and assayed aliquots for N-state protease activity as a function of time. The very small fraction of I that refolds during the incubation was detected with an assay of higher sensitivity than that used previously¹ (Fig. 2b). Based on the concentration of refolded N and the total protease concentration, the initial rate of folding k_f was found to be $1.18 \times 10^{-11} \pm 0.06 \times 10^{-11} \text{ s}^{-1}$ at 4 °C ($t_{1/2} > 1,800 \text{ y}$) and $6.0 \times 10^{-12} \pm 0.4 \times 10^{-12} \text{ s}^{-1}$ at 25 °C ($t_{1/2} > 3,600 \text{ y}$). These rates are significantly slower than the rates of unfolding. The equilibrium free energy, calculated from the ratio of k_f to k_u , favours I by $4.0 \pm 0.1 \text{ kcal mol}^{-1}$ at 4 °C and by $5.9 \pm 0.5 \text{ kcal mol}^{-1}$ at 25 °C. Therefore, the I state of α LP, not the N state, is at the minimum free energy over a broad range of temperatures.

α LP folds efficiently to its metastable native conformation only with the assistance of its pro-region. The catalysed folding reaction has two time constants, with the population of the fast phase being

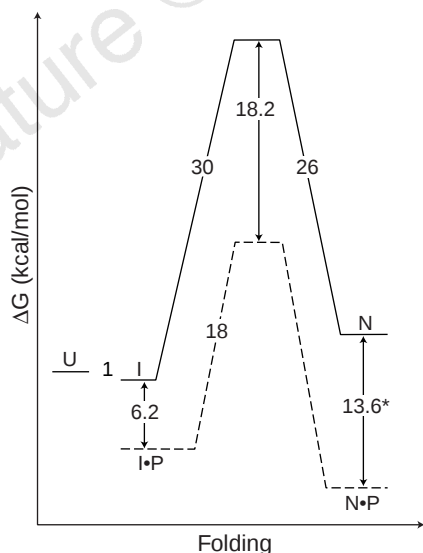


Figure 3 Free energy of α LP folding with and without its pro-region at 4 °C. Using the pro-region I-state binding constant (Fig. 2c), the pro-region N-state inhibition constant¹⁷ and the rates of folding and unfolding (Fig. 2a, b) in combination with transition state theory, the free-energy profile of α LP folding is drawn. Dashed lines denote energetic changes resulting from the addition of pro-region. The free-energy difference between I and U is not precisely known but is estimated to be less than $\sim 1 \text{ kcal mol}^{-1}$ (see Fig. 1b). Asterisk indicates that the pro-region binding of the native state was measured at 25 °C and pH 8.

roughly twice that of the slow phase. The slow phase, which is essentially independent of pro-region concentration, has a time constant of $\sim 200 \text{ s}$ (data not shown), and may be a consequence of proline isomerization, as α LP has one *cis* and three *trans* prolines. The fast phase rates are plotted as a function of pro-region concentration in Fig. 2c. From these data, the pro-region is calculated to bind I with $13 \pm 5 \mu\text{M}$ affinity and the folding rate is calculated to be $0.037 \pm 0.004 \text{ s}^{-1}$.

With the data shown in Figs 1 and 2, a free-energy diagram for the protease-folding reaction in the presence and absence of pro-region can be constructed (Fig. 3). Remarkably, both I and U are more stable than the folded N state. The pro-region shifts the energetics thermodynamically to favour the N-pro-region complex over the I-pro-region complex, thereby providing the driving force that brings the protease into its native, active conformation. Proteolytic degradation of the pro-region then 'locks' α LP in its metastable active N state. The pro-region is much more than a native-state template. It actively catalyses the folding reaction by stabilizing the folding transition state. The pro-region binds the folding transition state more tightly than I or N, and mutants of pro-region have been identified that preferentially alter this transition-state stabilization⁵.

Investigations into the enthalpic and entropic contributions to the free energy difference (ΔG) between I and N provide insight into the unusual stability of the unfolded states of α LP. Because direct measurement of the enthalpy difference (ΔH) between I and N is not possible, we measured ΔH by titration calorimetry using the thermodynamic cycle shown in Fig. 4. At 10 °C, the native conformation is favoured enthalpically by $18 \pm 1.5 \text{ kcal mol}^{-1}$ over I. Therefore, the stability of I must be entropic in origin.

Comparison of α LP with the homologous but thermodynamically stable protease, chymotrypsin, reveals an intriguing potential source of excess entropy. The α LP sequence contains 16% glycine compared to only 9% in chymotrypsin. Glycines lack a side chain, which increases the number of conformations accessible to unfolded states. The ten additional glycines in α LP are predicted to contribute an additional $\sim 7 \text{ kcal mol}^{-1}$ of configurational entropy to the unfolded state compared to that of chymotrypsin at 4 °C (ref. 6). Removal of this entropic source alone would be sufficient to place the N state at the global free-energy minimum for α LP.

An unusually low conformational entropy of the α LP N state may also contribute to increasing the unfolding entropy. Although native states are often dynamic, the N state of α LP is quite rigid: the protease is not readily digested by itself or by other proteases, it has ~ 40 core amides with $>10^{10}$ protection factors, and the B factors are unusually low (J. Davis, J.L.S. and D.A.A., unpublished results, and ref. 3). Some of the rigidity may arise from the fact that the

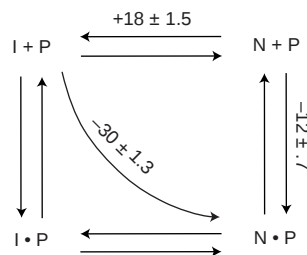


Figure 4 Dissection of the free energy difference between I and N. The enthalpy difference between I and N was determined at 10 °C using the outlined thermodynamic cycle. The enthalpies of I + P to N + P and N + P to N * P were measured. The total enthalpy change around the cycle must equal zero. Therefore, the enthalpy of the N- to I-state transition is $+18 \pm 1.5 \text{ kcal mol}^{-1}$ at 10 °C.

loops in α LP are generally shorter and are likely to be less flexible than those found in chymotrypsin. Many of the glycines unique to α LP are also found in these loops and so may reduce the entropy of the N state while increasing the configurational entropy of the I state.

The rigidity of the native state of α LP may be the product of evolutionary pressures to suppress autolysis and thereby extend the lifetime of the protease, which provides nutrients for its host. The unfavourable and presumably highly cooperative unfolding barrier of α LP would be expected to reduce sources of proteolytic sensitivity such as local breathing motions and/or partial unfolding of the native state. Similarly large kinetic barriers are likely to be present in most extracellular bacterial proteases; virtually all are synthesized with pro-regions and, where examined, the pro-regions catalyse protease folding⁷⁻⁹. There is a strong correlation between high glycine content and the presence of a pro-region. As the average glycine content of α LP and other chymotrypsin-like proteases made with pro-regions is 18%, compared with 9% in family members that do not contain a pro-region, incorporating more glycines may be one mechanism by which large kinetic barriers and pro-region catalysts have co-evolved in extracellular bacterial proteases in order to prolong their functional lifetime.

Large barriers have been observed in the folding of several other proteins that are not synthesized with pro-regions. The kinetic barriers of haemagglutinin, luciferase, the serpin PAI-1, and of the prion protein PrP, all function to enhance the stability of one compact native-like state relative to another¹⁰⁻¹³. In contrast, α LP has evolved a kinetic barrier to isolate its native state from unfolded states. This has enabled the protease to optimize its functional properties independently of the thermodynamic consequences, resulting in a kinetically stable but thermodynamically unstable native state. □

Methods

Protein expression and purification. α LP and pro-region were prepared according to published protocols¹⁴. The construct for the inactive mutant (S195A) has been described¹⁴. Expression and purification of S195A was as described¹⁵, except that the medium contained 1.5% yeast extract, 1% NaCl, 60 mM ACES, pH 6.3, and induction was at 12 °C. Pro-region was digested from the non-covalent S195A-pro-region complex with trypsin-coupled beads at pH 7 or with pepsin at pH 3. The S195A X-ray structure is virtually identical to that of wild-type protease (S.S.J., S. Rader and D.A.A., unpublished results).

Analytical ultracentrifugation. Sedimentation equilibrium experiments at 4 °C were carried out on I state in 10 mM potassium acetate, pH 5, on a Beckman Model XLA analytical ultracentrifuge at speeds of 18,000 and 22,000 r.p.m., scanning at 230 nm and 280 nm. Data were evaluated using XL-A data analysis software and fitted best at all concentrations to the single non-associating species model. Global analysis of nine data sets gave a relative molecular mass (M_r) of 21.7K with a 95% c.i. of ± 0.9 K.

Urea denaturation. The stability of I state was measured as a function of urea at 225 nm in an AVIV 62DS circular dichroism spectropolarimeter at 4 °C. 3.5 μ M I state in 10 mM potassium acetate at pH 5 was titrated with 4 M and 8 M urea. The unfolded baseline was calculated by a linear fit of the 3–6 M urea data.

Wild-type protease unfolding. 4 μ M protease in 0.98 M guanidinium-HCl, 10 mM potassium acetate, pH 5, was incubated at 25 °C. Activity was measured 13 times on duplicate aliquots according to published protocols over 80 days¹.

S195A unfolding. Reactions contained S195A in 10 mM potassium acetate, pH 5, at concentrations of 0.1–1.75 μ M (fluorescence) or 7 μ M (circular dichroism). Fluorescence measurements were made in a 8100SLM-Aminco fluorimeter. Circular dichroism spectra were monitored on a Jasco J-710 spectropolarimeter. The error in the 25 °C rate constant was determined from the difference between the linear extrapolation and denaturant binding model¹⁶ fits.

Refolding from the I state to the N state. 2.4–6.5 μ M I state in 10 mM potassium acetate, pH 5, was incubated at 4 °C and 25 °C. Initially and at each timepoint, 3-ml aliquots of I state were treated with 50 μ g pepsin at pH 2.5, which digests the I state but leaves native protease intact, then concentrated. The protease was assayed by following the absorbance change at 324 nm in a solution containing 1 mM succinyl-Ala-Ala-Pro-Ala-thiobenzyl ester (Enzymes Systems Products), 250 μ M aldrithiol-4 (Sigma), 2.5% DMSO and 0.1M Tris, pH8. Addition of 1 μ M pro-region inhibited the observed activity, indicating that activity is due to α LP N state. The concentration of native protease, [N], was determined from a standard curve. The detection limit is ~ 6.0 fmolar. Intact I state, [I]₀, was measured at each timepoint, as it decreases as a function of time owing to I-state proteolysis by refolded N.

Data from three experiments were fitted simultaneously to determine k_f using data in which the amount of intact I state was within 10% of the starting value. Calculations are based on a reversible equilibrium, $I \leftrightarrow N$. The rate equation for this equilibrium is $[I]_t/[I]_0 = [1/(k_f + k_u)] \times [k_f \{ \exp - (k_f + k_u)t \} + k_u]$, where the concentration of I state at time t is given by $[I]_t = [I]_0 - [N]$, assuming $[I]_t = [I]_0$ at $t = 0$. As both $k_f t$ and $k_u t$ are small, the Taylor series approximation is applied, yielding: $k_f = ([N]/[I]_0)/t$.

Refolding in the presence of the pro-region. Pro-region-catalysed refolding of the I state was measured at 4 °C in 20 mM potassium phosphate at pH 7.2 (ref. 1). The pro-region:I state concentration ratio was maintained at $\geq 15:1$ and reaction rates were fitted by a five-parameter double exponential. The plot of the fast phase rates as a function of pro-region concentration was analysed by assuming a fast $I + P \leftrightarrow I \cdot P$ pre-equilibrium followed by a slow, concentration-independent $I \cdot P \rightarrow N \cdot P$ transition.

Titration calorimetry. Enthalpies were measured in 20 mM potassium acetate, 42 mM guanidinium-HCl, pH 5, buffer using a Microcal Omega titration calorimeter. Pro-region (~ 100 μ M) was titrated with I or N from 8–12 °C. The enthalpies at 10 °C are the average of nine measurements. Temperatures below 8 °C and above 12 °C were not used owing to low signal-to-noise and complications from pro-region unfolding, respectively.

Received 15 May; accepted 30 July 1998.

- Baker, D., Sohl, J. L. & Agard, D. A. A protein-folding reaction under kinetic control. *Nature* **356**, 263–265 (1992).
- Silen, J. L. & Agard, D. A. The α -lytic protease pro-region does not require a physical linkage to activate the protease domain *in vivo*. *Nature* **341**, 462–464 (1989).
- Fujinaga, M., Delbaere, L. T., Brayer, G. D. & James, M. N. Refined structure of α -lytic protease at 1.7 Å resolution. Analysis of hydrogen bonding and solvent structure. *J. Mol. Biol.* **184**, 479–502 (1985).
- Makhatadze, G. I. & Privalov, P. L. Energetics of protein structure. *Adv. Protein Chem.* **47**, 307–425 (1995).
- Peters, R. J. *et al.* Pro region C-terminus: Protease active site interactions are critical in catalyzing the folding of α -lytic protease. *Biochemistry* **37**, 12058–12067 (1998).
- D'Aquino, J. A. *et al.* The magnitude of the backbone conformational entropy change in protein folding. *Proteins* **25**, 143–156 (1996).
- Baker, D., Shiao, A. K. & Agard, D. A. The role of pro regions in protein folding. *Curr. Opin. in Cell Biol.* **5**, 966–970 (1993).
- Eder, J. & Fersht, A. R. Pro-sequence-assisted protein folding. *Mol. Microbiol.* **16**, 609–614 (1995).
- Bryan, P. *et al.* Catalysis of a protein folding reaction: mechanistic implications of the 2.0 Å structure of the subtilisin-prodomain complex. *Biochemistry* **34**, 10310–10318 (1995).
- Clark, A. C. *et al.* Kinetic mechanism of luciferase subunit folding and assembly. *Biochemistry* **36**, 1891–1899 (1997).
- Chen, J. *et al.* A soluble domain of the membrane-anchoring chain of influenza virus hemagglutinin (HA2) folds in *Escherichia coli* into the low-pH-induced conformation. *Proc. Natl Acad. Sci. USA* **92**, 12205–12209 (1995).
- Harrison, P. M., Bamborough, P., Daggett, V., Prusiner, S. B. & Cohen, F. E. The prion folding problem. *Curr. Opin. Struct. Biol.* **7**, 53–59 (1997).
- Wang, Z., Mottonen, J. & Goldsmith, E. J. Kinetically controlled folding of the serpin plasminogen activator inhibitor 1. *Biochemistry* **35**, 16443–16448 (1996).
- Sohl, J. L., Shiao, A. K., Rader, S. D., Wilk, B. J. & Agard, D. A. Inhibition of α -lytic protease by pro region C-terminal steric occlusion of the active site. *Biochemistry* **36**, 3894–3902 (1997).
- Mace, J. E. & Agard, D. A. Kinetic and structural characterization of mutations of glycine 216 in α -lytic protease: a new target for engineering substrate specificity. *J. Mol. Biol.* **254**, 720–736 (1995).
- Myers, J. K., Pace, C. N. & Scholtz, J. M. Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci.* **4**, 2138–2148 (1995).
- Baker, D., Silen, J. L. & Agard, D. A. Protease pro region required for folding is a potent inhibitor of the mature enzyme. *Proteins* **12**, 339–344 (1992).

Acknowledgements. We thank P. Walter and P. Peluso for use of their fluorimeter, S. Marqusee for use of her Aviv CD and titrator, Y.-R. Yang for assistance with analysis of data from analytical ultracentrifugation, N. Sauter for assistance with figures, and A. Derman, M. Gustafsson, T. Mau, A. Shiao and H. S. Chan for comments on the manuscript. This work was supported by the Howard Hughes Medical Institute. S.S.J. is supported by a Howard Hughes Medical Institute predoctoral fellowship. J.L.S. was supported by NIH training grants.

Correspondence and requests for materials should be addressed to D.A.A. (e-mail: agard@msg.ucsf.edu).

Programming predominates

This roundup includes bioinformatics, image analysis and frame grabbing tools, bibliographic and reference management systems and a package that turns a variety of common programs into data acquisition systems.

Unscrambler

From Camo

www.camo.no

A tool for multivariate modelling, prediction, classification and experimental design

The program is divided in two parts: the 'editor', which is a spreadsheet-like data table where data can be displayed numerically, as well as transformed, and the 'viewer', which provides a graphical look at data and results. The software is said to be well suited for such applications as spectroscopy data calibration, data exploration, quality analysis, experimental design, environmental analysis, prospecting, and process and sensory analysis. The current version is written for the Windows 95 and Windows NT operating systems. The company can be found on the Internet at www.camo.no.

Reader Enquiry No. 100

Foundation Library

From NAG

www.nag.co.uk

A strict subset of the full NAG Fortran Library for users who do not need the depth of the full library, with its 1,000 user-callable routines

The routines are organized into chapters; each chapter is devoted to a branch of numerical or statistical computation, with a full list of the routines available from NAG. Routines include minimization, differential equations, partial differential equations, Fourier transforms, quadrature, linear algebra and linear least-squares problems (dense or sparse).

Reader Enquiry No. 101

Wedge Pro 32

From Vivcott Technology www.applegate.co.uk

This data acquisition package allows users to link any piece of equipment with a serial port to their favoured Windows software

For example, Microsoft Excel can be turned into a data logging program without having to do any programming in C or Visual Basic. WinWedge adds serial data acquisition and instrument control to Excel, Access, VB, MMIs, math and statistical software for any Windows, Windows95 or NT application. It automatically collects data from such laboratory instruments as pH meters, balances, moisture analysers, barcode readers, titrators, spectrometers, ion meters, radiation counters, barcode scanners, data loggers, modems. WinWedge performs real-time analysis, charting and graphing of your instrument or network data, says Vivcott.

Different instruments can simultaneously send data to different applications or to different 'fields' within the same application. In addition to data collection, the program can send data and commands out through serial ports to devices.

Reader Enquiry No. 102

Enhancing your image

UltraPlus

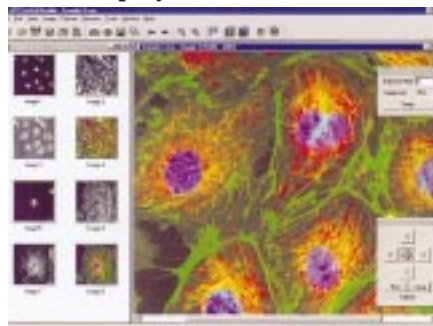
From LSR AstroCam

www.lsr.co.uk

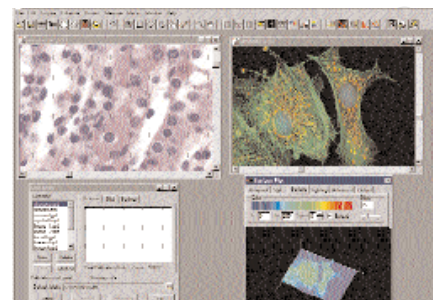
A new camera control and image analysis software interface for the company's Ultimate range of digital CCD cameras

This software is said to provide a high-performance, low-light-level image acquisition and analysis interface for the UltraPix range of 12- to 16-bit cooled CCD cameras. The UltraPlus Windows95 camera interface comprises user-friendly camera control and spatial modules. These interactive modules are said to provide a flexible working environment for both static and dynamic fluorescence, luminescence and brightfield imaging. The spatial module is used for grabbing multiple wavelength images or groups of images, which are then displayed in a thumbnail gallery. Morphometric measurements and densitometric quantitative analysis are then applied to up to 100 user-defined regions within a chosen image. The camera control module allows on-chip, asymmetric binning and sub-array selection, making it well suited for high-speed image acquisition at up to 1,500 images per second. Automated, time-lapse studies, detailed image statistics and on-line diagnostics are also provided. The program provides a variety of image display and analysis techniques, as well as the tools to present the best results possible. Images are processed using background subtraction, flat-field correction and thresholding to provide enhanced brightness, contrast and sharpness.

Reader Enquiry No. 103



UltraPlus for camera control and image analysis.



Seeing in stereo with ImagePro Plus.

Image-Pro Plus

From Media Cybernetics www.mediacy.com

Designed to increase productivity in microscopy and image acquisition

Image-Pro Plus includes a full range of counting, sizing, statistical and image enhancement tools. Version 4.0 includes new features for life science research, such as a caliper tool, a grid-making toolkit and a surface plot tool for three-dimensional image intensity. The new caliper tool allows researchers to locate edges along straight or irregular lines and circles automatically, and to detect and measure distances between edges. The grid-making tool is designed for researchers currently using stereological methods for estimating particle number, area, volume and structural relations. The tool allows these users to make and save their own grids. With just the click of a mouse, the surface plot tool displays image intensity in a full colour three-dimensional plot, allowing users to rotate, tilt, colourize or wireframe the plot.

Reader Enquiry No. 104

Quartz PCI System

From Nissei Sangyo America www.nissei.com

Version 4.2 is a full 32-bit application for microscopic image capture and management in Windows95 or NT

This program permits users to capture and manage digital images from such devices as SEMs, STEMs, TEMs, light microscopes, video sources, TWAIN devices and CD SEMs, all using a single family of software. Version 4.2 has an angle measurement tool that has been added, as well as improved colour image support with tools that adjust image colour, balance, hue, saturation and lightness. In addition, there is an improved histogram function that allows interactive adjustment of image contrast, brightness

and gamma. The new 'send to' function transfers documents directly from other applications to the PCI database for storage and retrieval in the same manner as images. There is also a 'process stereo' function that provides for stereo-pair alignment and, to help users cope with large volumes of images, there is a family of scalable ODBC-compliant database solutions.

Reader Enquiry No. 105

RatioPlus 2.1

From Scanalytics www.scanalytics.com

A set of IPLab Spectrum-compatible extensions that enable users to acquire, process and analyse dual-wavelength image data

For use in Ca^{2+} and pH ratio experiments with high-quality cameras and wavelength switching systems, version 2.1 has been designed for flexibility, ease of use and improved experimental consistency. Users can acquire image sequences that can be stored in memory or on a hard disk, generate plots showing changes in selected areas or cells through time and animate the resulting image sequence as a QuickTime movie. Mosaic images can also be produced quickly and easily for presentations. Moreover, the software offers flexibility to design customized experiments, as well as the flexibility to use a wide range of compatible camera types and wavelength switching equipment. A redesigned interface allows scientists to plot multi-frame objects, such as means, ion concentration and pH concentration over time. The program also makes it possible to perform real-time background subtraction during live experiments.

Reader Enquiry No. 106

Snapper

From Data Cell www.datacell.co.uk

The company's latest frame-grabbers include support for the Windows98, 95 and NT4.0

Windows support includes SnapperTool, a configuration utility that supports existing and new cameras through a graphical user interface. Depending on the nature of the camera and board (for example, line-scan or area-scan), the dialogues are reconfigured to show the specific controls that are relevant. For instance, there is a complete set of custom dialogues for Kodak Megaplug cameras, allowing complete control over integration times and shuttering — Hamamatsu support is due to be released later this year. SnapperTool was built using the ActiveX developers kit, which allows customized applications to offer the same powerful interface over the entire Snapper range. DataCell also provides TWAIN, Image-Pro Plus and LabView drivers built using the same ActiveX support.

Reader Enquiry No. 107

Bibliographics research

EndNote

From Niles Software www.niles.com

Niles and BioMedNet have collaborated to integrate Evaluated Medline and EndNote

Medline now includes a 'direct-to-Endnote' option, allowing users to transfer search results to their EndNote databases. This integration is said to eliminate the extra steps of downloading and importing. According to Niles, users can conduct a search on Evaluated Medline and then simply click a button to have the search results transferred to their Endnote database to create a bibliography instantly.

Reader Enquiry No. 108

Reference Manager 8.5

From Research Information Systems

www.risinc.com

Version 8.5 is now available for use within Windows95 and Windows NT

This bibliographic software now comes with spell checking, new database tools and advances Web functionality. Users can now spell check their references, and include custom dictionaries in a system that offers phonetic and typographic suggestions, identifies repeated words and supports multiple language dictionaries. Custom dictionaries allow term lists to be imported into a working dictionary. The new database tools provide property sheets, common dialogues, object linking and embedding, long file names, toolbars and right mouse support. Users may drag-and-drop throughout the program to move information within a reference, between databases or even to other applications, such as a word processor.

Reader Enquiry No. 109

Get A Ref

From Adept Scientific www.adeptsience.co.uk

On-line reference management and retrieval with access data across the Internet

This program allows users to locate references from a variety of different sources, format them and move them into a word processing document. It should prove useful to researchers who need access to remote material. The program will automatically import references from several on-line sources. In addition to the Medline, Silverplatter, CDPlus, Compact Cambridge and Data Star reference management systems, GetARef can now access and convert information from the UK's BIDS networked information for the higher education and research communities, which is run by the University of Bath. It can also now accept data from Papyrus, the widely used DOS-based system. Moreover, the latest version of the conversion module contains routines that allow

the conversion of data and complete files from Papyrus — these can then be dropped into the appropriate place within the users' word processing documents. A plain text converter has also been introduced for the first time, and other non-supported formats (including on-line data types) can be supplied on request. The converter, GarConv, is available free of charge for users at www.adeptsience.co.uk.

Reader Enquiry No. 110

These notes are compiled by Brendan Horton from information provided by the manufacturers. For more details, fill in the reader service card bound inside the journal.

ADVERTISEMENTS

NEWCASTLE UNIVERSITY
MOLECULAR BIOLOGY UNIT

PROTEIN SEQUENCING
Membrane-bound, solid or solution samples
£12 per cycle
5 cycles minimum - no set up charge

For more information please contact Dr. Jon Gray:
Phone/Fac: +44 191 222 8512
Email: JOE.GRAY@newcastle.ac.uk
DNA sequencing (ABI 377-8000, 877 Cooled)
also available at competitive prices

READER ENQUIRY NO. 42

HVS Water for Windows software runs under Windows 3.x, 95, 98 is now proven in the U.S. and worldwide. For (Movie) Water Histo. Also software for open field & radial-arm maze. System includes video camera, monitor, VT200 Advanced Tracker, software package.



HVS IMAGE PO Box 104, Hampton, UK
In the U.S. and Canada call 1-800-323-9261.
email@hvsimage.com www.hvsimage.com
Newcomer, LA, Nov 8-11, 1998. Booth 1237.

READER ENQUIRY NO. 52

Don't react! DNA with Wako's
DNA Extractor Kit
or isolate genomic DNA with
DNA Extractor WB Kit
Both use the Nal Method
No Phenol No Chloroform

Wako

Wako BioProducts
Wako Chemicals U.S.A., Inc.
1800 Belwood Road
Richmond, VA 23137
Telephone: (804) 871-7577
Fax: (804) 871-7710
(800) 962-6266
WakoProducts@wako.com

READER ENQUIRY NO. 39